Research and Innovation Action

Social Sciences & Humanities Open Cloud

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

Deliverable 7.3 Marketplace - Interoperability

Dissemination Level	PU
Due Date of Deliverable	31/12/2021 (M36)
Actual Submission Date	15/12/2021
Work Package	WP7 - Creating the SSH Open Marketplace
Task	Task 7.3 Marketplace Interoperability
Type	Report
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p.1 – p.33

Abstract:

Report summarising the findings and state of development regarding the interoperability of the SSH Open Marketplace with external systems. It delivers an overview of the dynamic broader technology landscape and details modes of interaction and experience gathered when interconnecting the various systems and the Marketplace.

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	06/2021	first draft skeleton	Dieter Van Uytvanck, Alexander König
0.1	07/2021	drafting	Dieter Van Uytvanck, Alexander König with contributions from Carsten Thiel, Frank Fischer, Laure Barbot, Edward Gray, Klaus Illmayer, Alireza Zarei, Matej Ďurčo.
0.2	11/2021	Version ready for peer review	Alexander König
0.3	11/2021	Peer review	Daan Broeder (CLARIN), Andrea Scharnhorst (DANS)
0.4	8/12/2021	Address peer review comments	Alexander König
1.0	13/12/2021	Submission	CESSDA

Author List

Organisation	Name	Contact Information
CLARIN	Alexander König	alex@clarin.eu
CLARIN	Dieter Van Uytvanck	dieter@clarin.eu
CESSDA	Carsten Thiel	carsten.thiel@cessda.eu
DARIAH	Frank Fischer	frank.fischer@dariah.eu
DARIAH	Laure Barbot	laure.barbot@dariah.eu
CNRS	Edward Gray	edward.gray@huma-num.fr
DARIAH/OEAW	Klaus Illmayer	Klaus.Illmayer@oeaw.ac.at
DARIAH/UGOE	Alireza Zarei	alireza.zarei@gwdg.de
DARIAH/OEAW	Matej Durco	Matej.Durco@oeaw.ac.at

Executive Summary

The SSH Open Marketplace is one of the key exploitable results of the SSHOC project. It is designed to be a central information hub where researchers can go to find detailed information about software tools, relevant publications, datasets and workflows. Within the Marketplace contextualisation should be provided, linking, for example, publications to the tools or datasets they mention.

This deliverable, D7.3 Marketplace Interoperability, focuses on the process of populating the SSH Open Marketplace, thereby complementing the deliverables on the technical implementation (D7.2 Marketplace Implementation) and the curation (D7.4 Marketplace Data Population and Curation) of the Marketplace.

In the beginning of the SSHOC project a list of potential sources that could be added to the new SSH Open Marketplace was compiled. During the first phase of the project, each source in this list was then investigated regarding its potential to be added to the Marketplace. This deliverable details how the selection process of potential sources worked. Each of the potential sources will be described briefly, it will be highlighted why they would be of interest to be added to the Marketplace and discussed whether the decision was made to ingest them into the Marketplace during the run-time of the SSHOC project.

The document is structured by the decision made for each specific source, starting with a section of all the sources that have already been ingested in the Marketplace by the time of writing. This is followed by the sources that were decided to be ingested in the course of the project and that are currently in the process of being ingested. Then the sources are listed that were considered but discarded for one reason or another (e.g. poor quality of metadata, not enough new, unique items) and finally there is a section with sources that would be great additions but cannot be easily ingested using the automatic ingestion pipelines that are currently available. These sources might later be considered for manual inclusion into the Marketplace. This document also provides an overview table showing for each source how many individual items have been or will likely be ingested into the Marketplace and whether the source will be ingested only once or is considered for continuous ingestion, because it is expected that there will be frequent updates at the source that should also be reflected in the Marketplace. At the end of the document there is also a section discussing possible future steps that could be taken to improve the quality of the Marketplace after the project has ended.

Abbreviations and Acronyms

ADHO	Alliance of Digital Humanities Organizations
API	Application Programming Interface
CESSDA	Consortium of European Social Science Data Archives
CLARIN	Common Language Resources and Technology Infrastructure
CMS	Content Management System
CSV	comma-separated values
D	deliverable
DACE	Data Aggregation and proCessing Engine
DANS	Data Archiving and Networked Services
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DH	Digital Humanities
DiRT	Directory for Research Tools
EAD	Encoded Archival Description
EGI	European Grid Infrastructure
ELG	European Language Grid
EOSC	European Open Science Cloud
ESS	European Social Survey
GLAM	Galleries, Libraries, Archives, and Museums
ICPSR	Inter-university Consortium for Political and Social Research
LRS	Language Resource Switchboard
NLP	Natural language processing
PH	The Programming Historian
PSNC	Poznan Supercomputing and Networking Center
SSH	Social Sciences and Humanities
SSHOC	Social Sciences and Humanities Open Cloud
SSK	Standardization Survival Kit
TAPoR	Text Analysis Portal for Research
TEI	Text Encoding Initiative
TERESAH	Tools E-Registry for E-Social science, Arts and Humanities
TRL	Technology Readiness Level
VLO	Virtual Language Observatory
WP	work package
XML	Extensible Markup Language

Table of Contents

1. Introduction

The SSH Open Marketplace is designed to be a central information hub where researchers can go to find detailed information about software tools, relevant publications, datasets and workflows. To be of use to the research community in the Social Sciences and Humanities, it is important to populate the Marketplace with a good selection of relevant entries while also ensuring not to overwhelm the users. This deliverable details the decision process in task 7.3 " Marketplace Interoperability" to select the best sources for the initial population of the Marketplace with relevant entries.

Considering the typology of items that were identified as relevant for the Marketplace, namely *datasets*, *tools and services*, *training materials*, *publications* and *workflows*, as a first step a list of possible sources for the marketplace was collected in a kind of crowdsourcing among the participants in this task and their networks. This first longlist was not filtered and simply contained any source that those participants in this task[1] could think of, that might possibly be worth including in the SSH Open Marketplace. The term "source" is being used here for a very heterogeneous list of entities containing information about the items of interest for the Marketplace, ranging from catalogues for data (e.g. the Virtual Language Observatory[2]) or tools (e.g. TAPoR[3]), where the items were harvested automatically or manually collected, to lists that were compiled during the SSHOC project (e.g. SSHOC service catalogue[4]) or publications from relevant conference series (e.g. DH conference papers[5]). The size of the sources also varies a lot, from catalogues containing thousands or millions of items to small hand-curated lists that only contain dozens or even a handful of items. As one of the goals for the Marketplace was improving usability by not overwhelming users with thousands of results for any given search, it was decided not to ingest all items from the very large sources, i.e. data catalogues like the Virtual Language Observatory, but on the other hand, these (data) catalogues are very useful services for SSH researchers and therefore it was decided to include these sources as a single entry in the Marketplace. Additionally, these entries were tagged with the keyword "data catalogue" to group them together among the "tools & services" Marketplace category.

In the next step, T7.3 members looked at each of these sources (or rather, the items contained therein) and made a decision on whether it should be included in the Marketplace or not. The most important criteria for this decision were

[1] The following partners of Task 7.3 have been involved in the creation of this list: CLARIN ERIC, DARIAH/PSNC, CNRS, DARIAH/OEAW, DARIAH/UGOE, SWC, CLARIN/Athena, CNR. All SSHOC Work Package 7 members were consulted as well, and contributions from the community were also collected as part of a consultation platform set up by SSHOC Work Package2, in order to guarantee the diversity of disciplines and background.

[2] Virtual Language Observatory website: https://vlo.clarin.eu [13.12.2021]

[3] Text Analysis Portal for Research (TAPoR) website: Text Analysis Portal for Research [13.12.2021]

[4] SSHOC service catalogue: https://www.sshopencloud.eu/service-catalogue [13.12.2021]

[5] Abstracts of the DH Conference on dblp website: https://dblp.org/db/conf/dihu/index.html [13.12.2021]

- the **quality of the (meta)data:** higher quality data will increase the quality of the Marketplace; this means that the more of the metadata fields in the Marketplace data model are covered by an item, the better (items with just a title and a link are inferior to ones containing a long description and maybe even a list of contributors) and the more items within the source have good quality metadata , the higher the overall quality of the source; as a corollary of this, it was decided to focus on primary sources and exclude sources that only aggregate information that can be found elsewhere; such "2nd hand metadata" introduces additional points of failure and using the primary source instead will likely result in higher quality metadata
- the **uniqueness of the data:** if another source covers the same items already, there is no need to get them again from another source; even though duplicate entries are dealt with within the Marketplace[6], ingesting identical (or worse almost identical) entries multiple times uses up valuable human resources that will be better put to use elsewhere considering the limited timeframe and personnel of the project
- the **technical interface:** it is easier to import a source into the Marketplace if it offers a well-documented API; considering the limited human resources of the project, sources which needed a lot of additional work before they could be fed into the ingestion pipeline were placed very low on the priority list
- the **expected usefulness** for the SSH community: are the items in this source possibly of use to SSH researchers; it could be that they are widely used already or that they offer new and interesting possibilities
- the **representativity** of the source for the SSH domain: with the various partners in the SSHOC project coming from different parts of the wide field of SSH (e.g. CESSDA being mostly concerned with social sciences, while CLARIN is more linguistics/language focused), considerable effort was made to find sources that represent all the different facets of the SSH landscape; it also made sense to prioritize sources coming from the SSHOC project itself, assuming that they will have this kind of representativity baked in already.

Apart from these criteria, **contextualization** of entries in the Marketplace was a big focus. The idea being that by linking items together and showing the user relations between, for example, a publication and the software tools and datasets that were used in the research described therein, the usefulness of the Marketplace will increase a lot.

In section 2, all the sources that were considered are listed and it is explained for each of them why they were included in the Marketplace or why it was decided not to include them. The section begins with a table listing all the sources.

[6] Duplicate entries are identified after ingestion through a semi-automatic process using Python notebooks. This process can discover similar or identical entries and flag them for the system to create a suggested merged entry that the curation team needs to review before it becomes an accepted new entry in the Marketplace.

After the sources were selected, the actual ingestion process was done using two different pipelines, Poolparty (a pre-existing tool provided during the project by one of the project partners, Semantic Web Company)[7] and DACE (a tool that was newly developed by DARIAH/PSNC within this project)[8]. The idea was to use an existing tool to speed up the first ingestions. The use of Poolparty was free of charge during the project, because it was provided by one of the project partners. However, using this pipeline after the end of the SSHOC project would incur high costs. Therefore it was decided to develop a new tool in the course of the project that could take over this functionality and be available freely after the end of the project. For more details on the ingestion pipelines, including technical details, please see *D7.2 Marketplace Implementation*.

For each source it was decided whether it should be only ingested once or whether the source was very active, would change in time and so new versions of it should be ingested continuously. In the latter case, the idea was that the source will be re-ingested into the Marketplace at regular intervals, so that newly added items in the source will also be visible in the Marketplace, items that no longer exist can be deprecated in the Marketplace and any update of an item's metadata will be picked up by the Marketplace as well.

The ingestion workflow for each source started with mapping the source's metadata onto the Marketplace data model[9]. This mapping was then used by the ingestion pipeline to import the source into the Marketplace. Afterwards the imported items were checked automatically and manually to ensure that the mapping worked well. This resulted in an updated and improved mapping which then was used to re-ingest the source again.

This deliverable will focus on the selection process of the sources for the Marketplace and leave out all the technical details about the mapping of metadata and the actual ingestion process. To get the full picture see also the complementary deliverables *D7.2 Marketplace Implementation*[10] on the technical implementation details and *D7.4 Marketplace - Data population & curation*[11] on the curation of the Marketplace contents.

[7] PoolParty Semantic Suite: https://www.poolparty.biz/ [26.10.2021]

[8] DACE - Data Aggregation and proCessing Engine: https://gitlab.pcss.pl/dl-team/aggregation/dace [26.10.2021]

[9] See SSHOC *D7.1 System Specification - SSH Open Marketplace* for a first version of the Marketplace data model: https://doi.org/10.5281/zenodo.3547648 [26.10.2021]. Version 1.5 is described in SSHOC *D7.2 Marketplace Implementation*: https://doi.org/10.5281/zenodo.5749464 [10.12.2021].

[10] Matej Ďurčo, Laure Barbot, Klaus Illmayer, Sotiris Karampatakis, Frank Fischer, Yoann Moranville, Joshua Tetteh Ocansey, Stefan Probst, Michał Kozak, Stefan Buddenbohm, & Seung-Bin Yim. (2021). 7.2 Marketplace – Implementation (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5749465 [10.12.2021]

[11] Edward Gray, Nicolas Larrousse, Clara Petitfils, Laure Barbot, Frank Fischer, Matej Ďurčo, Klaus Illmayer, Cesare Condordia, Alexander Konig, Dieter Van Uytvanck, & Stefan Buddenbohm. (2021). D7.4 Marketplace – Data population & curation (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5783358 [10.12.2021]

2. Overview of external data sources

The following table contains all sources described in this section. They are listed in the table in the same order they appear in the text. For each source, the table lists the current status of the source at the time of writing, and for those that have been or are planned to be ingested, the expected number of items to be ingested into the Marketplace and the decision whether a source will be ingested once or continuously. In rows 2-5 sources are listed that have been ingested via the EOSC Portal Marketplace (see below for details) which is why they do not have any information in columns 2-4. The status "later" marks sources that have not been discarded, but are also not planned to be ingested during the course of the SSHOC project. This status means that the number of potentially ingested items and the type of ingestion are mostl not known yet for these sources.

Name of source	current status	estimated number of items to be ingested	continuous ingest or one-time
EOSC Portal Marketplace	ingested	25	continuous
- CESSDA Data Catalogue	N/A	N/A	N/A
- EGI Marketplace	N/A	N/A	N/A
- LINDAT/CLARIAH-CZ Services	N/A	N/A	N/A
- Virtual Language Observatory	N/A	N/A	N/A
DH conference papers via dblp	ingested	2,800	continuous
Humanities Data	ingested	300	one-time
Language Resource Switchboard (LRS)	ingested	55	continuous
The Programming Historian (PH)	ingested	170	continuous
Standardization Survival Kit (SSK)	ingested	400	one-time
TAPoR	ingested	1,400	continuous
CESSDA training resources	planned	100	continuous
CLARIN Resource Families	in-process	1,100	continuous
DARIAH-Campus	in-process	70	continuous

DARIAH contribution tool	planned	60	continuous
SSHOC Service Catalogue	in-process	20	one-time
SSHOC training materials	planned	50	one-time
SSH Conversion Hub	planned	50	one-time
SSH Training Discovery Toolkit	planned	300	one-time
Awesome Digital History	discarded	N/A	N/A
B2FIND	discarded	N/A	N/A
DARIAH-DE Collection Registry	discarded	N/A	N/A
ICPSR	discarded	N/A	N/A
METHODICA	discarded	N/A	N/A
TERESAH	discarded	N/A	N/A
DH Toychest	later	tbd	tbd
European Language Grid	later	470	tbd
GLAM Workbench	later	tbd	tbd
Innovations in Scholarly Communication	later	tbd	tbd
OpenMethods	later	tbd	tbd
Open Source Tools for Social Science Researchers	later	tbd	tbd
SERISS tools	later	tbd	tbd
Tools, data or services mentioned during T7.1 interviews	later	tbd	tbd
OpenAIRE	later	22,000	continuous

Table 1: SSH Open Marketplace sources overview

2.1 Sources ingested into the SSH Open Marketplace

This section provides an overview of all sources that were ingested into the marketplace by the time of writing. This selection contains to a large extent sources that were included in the beta release of the Marketplace, in December 2020[12]. Further ingests were done on a separate non-public staging instance of the Marketplace and happened in parallel to a high degree. This approach results in all of the remaining sources (see section 2.2) being ingested into the production instance of the Marketplace at once close to the end of the project in April 2022.

Each of the sources is described with at most a few paragraphs. It was decided to leave out the more technical details of the implementation, e.g. the mapping of metadata fields into the Marketplace data model. For implementation details see deliverable *D7.2 Marketplace Implementation*.

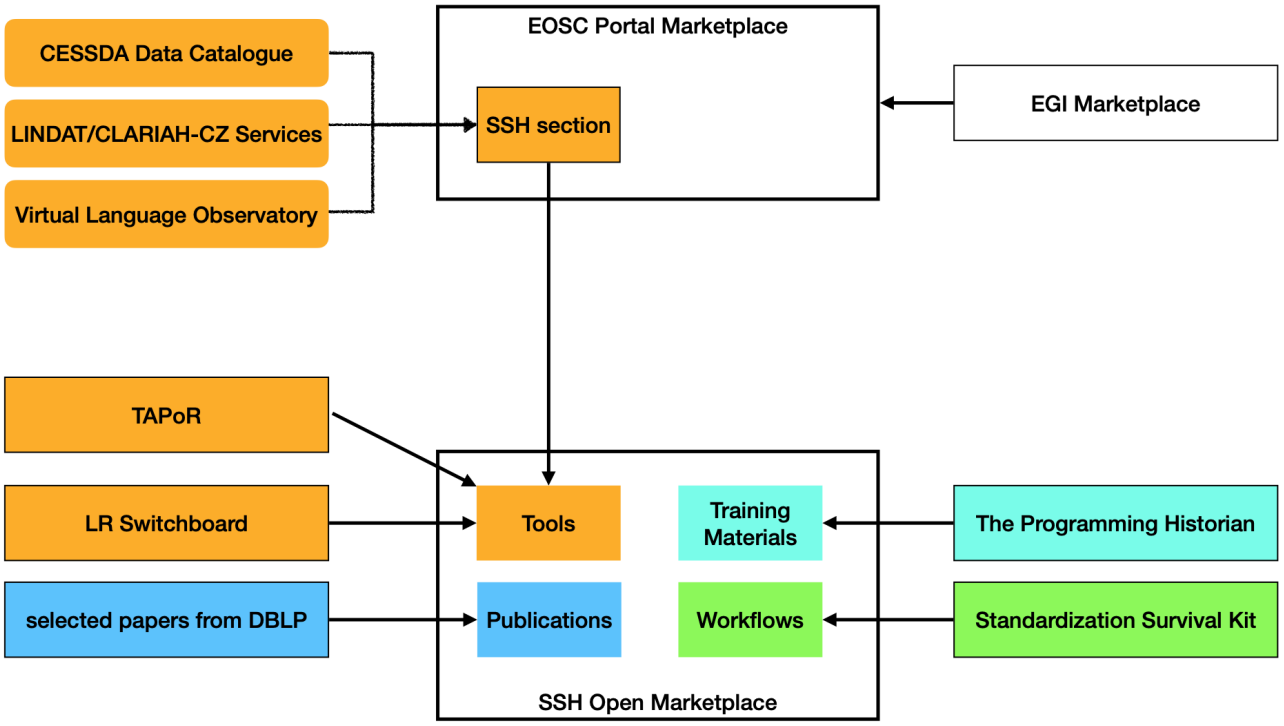

Figure 1: Flowchart of the currently included sources into the SSH Open Marketplace.
Rounded corners are used for manually created single entries within a marketplace. The colours signify the different types of items.

[12] See SSHOC Milestone 43 report for more details: Laure Barbot, Frank Fischer, Klaus Illmayer, Matej Ďurčo, Alexander König, Dieter Van Uytvanck, & Nicolas Larrousse. (2020). MS.43 -Marketplace - beta release (1.0). Zenodo. https://doi.org/10.5281/zenodo.4785194 [10.12.2021]

EOSC Portal marketplace

The EOSC Portal is a collaboration between the eInfraCentral[13], EOSCpilot[14], EOSC-hub[15], OpenAIRE-Advance[16], EOSCsecretariat.eu[17], EOSC Enhance[18] and the EOSC Future[19] project. The EOSC Portal Marketplace[20] was launched in November 2018 and is intended to provide an overarching catalogue of resources and services available through the European Open Science Cloud. It currently contains over 300 entries.

Since its inception, a lot of functionality has been added to enable EOSC resource and service providers to edit the metadata of their entries through the EOSC Portal. At the same time, many of the SSHOC partners (e.g. CESSDA, CLARIN and DARIAH) invested a lot of time to register their services with the EOSC Portal marketplace, since there are many details required to register an entry and since in the beginning the pioneering status of the EOSC Portal Marketplace required many curation efforts.

To avoid double work, some form of automated synchronisation between both marketplaces is desirable. It was decided to implement a **unidirectional** and continuous import from the EOSC Portal Marketplace – for those entries listed in the Humanities and Social Sciences sections – into the SSH Open Marketplace[21]. The choice for such a unidirectional synchronisation is motivated by the following reasons:

- The EOSC Portal Marketplace existed before the SSH Open Marketplace.
- The process of entering metadata into the EOSC Portal Marketplace is more demanding – there are more obligatory fields .
- Many of the required fields for the EOSC Portal Marketplace are not available in the SSH Open Marketplace, as they are not considered relevant for the SSH community (for example, the security contact email or the order type and payment model fields have been left out).
- There are stricter inclusion criteria[22] for the EOSC Portal Marketplace, such as a minimum Technology Readiness Level (TRL) of 7. Most entries in the SSH Open Marketplace do not contain this kind of information and it would require quite some additional effort to filter out

[13] European E-Infrastructure Services Gateway: https://cordis.europa.eu/project/id/731049; [30.08.2021]
[14] EOSCpilot: https://eoscpilot.eu/; [30.08.2021]
[15] EOSC-hub: https://www.eosc-hub.eu/; [30.08.2021]
[16] OpenAIRE website: https://www.openaire.eu/; [30.08.2021]
[17] EOSCsecretariat.eu: https://www.eoscsecretariat.eu/ [10.09.2021]
[18] EOSC Enhance page on EOSC portal: https://www.eosc-portal.eu/enhance [10.09.2021]
[19] EOSC Future website: https://eoscfuture.eu [10.09.2021]
[20] EOSC Portal Catalogue and Marketplace: https://marketplace.eosc-portal.eu/ [10.09.2021]
[21] See https://marketplace.sshopencloud.eu/search?categories=tool-or-service&order=label&f.source=EOSC+Marketplace
[22] Information about inclusion criteria in the EOSC Portal Marketplace FAQ: https://marketplace.eosc-portal.eu/help [10.09.2021]

only entries that could be synced back to the EOSC Portal Marketplace. Additionally, this would likely not be a lot of entries as a result.

If SSH service providers would like to be listed in both Marketplaces, the recommendation is to follow the onboarding procedure[23] for the EOSC Portal Marketplace.

As a result of this approach, the EOSC Portal Marketplace ended up already containing some of the sources that were considered for inclusion in the SSH Open Marketplace. These are briefly listed in the next section.

2.1.1 Resources included via the EOSC Portal Marketplace

CESSDA Data Catalogue

The CESSDA Data Catalogue[24] contains the metadata of all data in the holdings of CESSDA service providers. It is a one-stop-shop for search and discovery, enabling effective access to European research data for researchers. Details of over 30.000 data collections are listed. These are harvested from fifteen different CESSDA Service Providers. Given the size and type of data, it was decided to include the CESSDA Data Catalogue as a single entry, directly harvested from the EOSC Portal.

From 2022, the CESSDA Data Catalogue will be harvestable by all EOSC data portals using the OAI-PMH protocol, in accordance with the specifications for aggregated metadata provenance, ensuring proper deduplication. This will ensure the inclusion of Social Science data in the EOSC data landscape.

EGI Marketplace

All items from the EGI Marketplace[25], a special portal that collects services provided by members of the EGI federation, have also been included into the EOSC Portal Marketplace. Since there is an import chain in place from the latter to the SSH Open Marketplace, all relevant entries for Humanities and Social Sciences services coming from the EGI marketplace (currently none) will automatically be included.

LINDAT/CLARIAH-CZ Services

Many of the LINDAT/CLARIAH-CZ Services, including the Machine Translation and NLP pipeline used in the context of SSHOC Work Packages 4 and 3, have been included[26] into the EOSC Portal Marketplace. Via the automated synchronisation, they appear in the SSH Open Marketplace as well.

[23] Information for providers to the EOSC Portal: https://eosc-portal.eu/providers-documentation [10.09.2021]
[24] CESSDA Data Catalogue: https://datacatalogue.cessda.eu/; [19.07.2021];
[25] EGI Marketplace portal: https://marketplace.egi.eu/; [09.08.2021]
[26] Overview of LINDAT/CLARIAH-CZ resources in the Marketplace: https://marketplace.eosc-portal.eu/services?providers%5B%5D=189; [30.08.2021]

Virtual Language Observatory

The metadata catalogue of CLARIN, the Virtual Language Observatory[27], provides a unified search portal for all CLARIN repositories. Given its size (over a million records) and nature (mostly language data), it was decided to include the VLO as a single entry. Since its description was already entered in the EOSC Portal Marketplace, it was automatically imported from there.

2.1.2 Sources included in the beta release of the SSH Open Marketplace

DH conference papers via DBLP

The annual conference of the Alliance of Digital Humanities Organizations (ADHO) is the most important event in the Digital Humanities. The first conference took place in 1989. Currently every conference attracts around 1.000 participants, hundreds of peer-reviewed papers and posters are presented. **DBLP**[28] has been used as a broker to obtain the full texts of all conference papers since 2010, which were text-mined for mentioned tools to provide contextualisation. See deliverable *D7.2 Marketplace Implementation* and specifically subsection 4.6 "Extraction module" for more details on this.

Humanities Data

The Humanities Data[29] website collects and presents datasets and recipes stemming from Digital Humanities projects. The project is led by a single scholar, Matthew J. Lavin from Denison University in the United States. Metadata of 303 datasets and 16 recipes had been gathered at the time of the first data ingest to the SSH Open Marketplace. Only datasets were selected for ingestion since the featured recipes seemed too outdated and did not align with existing training materials or workflows in the Marketplace.

Despite having only limited metadata available for each dataset, Humanities Data has been considered a good source to initiate the dataset population of the SSH Open Marketplace with Humanities content. The Humanities Data metadata schema reuses data fields associated with Project Open Data version 1.1[30]. For all the datasets metadata ingested in the SSH Open Marketplace, at least a title, a description, contributors' names and a URL pointing to the dataset (mostly hosted on a Dataverse instance or at arXiv) have been retrieved.

[27] Virtual Language Observatory: https://vlo.clarin.eu; [30.08.2021]
[28] dblp: computer science bibliography: https://dblp.org/ [20.10.2021]
[29] Humanities Data website: https://humanitiesdata.com/; [19.07.2021]
[30] See DCAT-US Schema v1.1 (Project Open Data Metadata Schema): https://resources.data.gov/resources/dcat-us/ [20.07.2021]

Language Resource Switchboard

The Language Resource Switchboard[31] is a tool that helps users to find a matching language processing web application for their data. The tool was originally developed by CLARIN and has also been worked on in the context of the SSHOC project (see Deliverable D3.8)[32]. It is configured with a wide range of tools and webservices available mostly from the CLARIN community that can be invoked to directly process a certain file, either by uploading it to the switchboard or by providing a URL to a file that is available online. The switchboard automatically analyses the input file and offers only those tools and services that will work with the file's format (e.g. plain text or XML) and the text language. The user can also override these automatically detected values if necessary. As the Switchboard is basically a collection of well-working[33] and useful linguistic services and tools it was decided to add each of these tools as separate items to the SSH Open Marketplace. All tools and services integrated in the Switchboard can be queried via an API which was used to map this source into the Marketplace.

The Programming Historian

The Programming Historian[34] is an open-access journal that presents peer-reviewed tutorials in the digital humanities and digital history methodology. Programming Historian is available in four languages: English, Spanish, Portuguese and French. It is one of the most recognized open source journals in the digital humanities, and has won multiple awards for its content. The Programming Historian has 163 individual "lessons" (86 in English, 50 in Spanish, 17 in French, and 10 in Portuguese). The SSH Open Marketplace has ingested the lessons in English, with plans to ingest recipes in other languages in the future (see section 3 for a more detailed discussion of this).

The Standardization Survival Kit

The Standardization Survival Kit (SSK)[35] is an outcome of the Horizon 2020 project PARTHENOS[36]. It lists use cases from different research communities - mainly from the humanities - describing the application of digital methods and recommending the use of community-agreed standards. These use cases are called "scenarios" in the SSK and they consist of smaller "steps", explaining an action in more detail. Scenarios in the SSK are either generic ones - describing an ideal workflow - or they are derived from project specific approaches. To give some examples for scenarios: one showcases a workflow for digitising textual material, another one explains how to do linguistic annotations of corpora. Scenario

[31] Language Resource Switchboard interface: https://switchboard.clarin.eu/; [09.08.2021]

[32] Daan Broeder, Willem Elbers, Stefan Buddenbohm, Wolfgang Schmidle, Emanuel Dima, Matej Durco, Cesare Concordia, Maurizio Sanesi, & Emiliano Degl'Innocenti. (2021). D3.8 Implementation report and available SSHOC Switchboard and VCR services (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5608542 [13.12.2021]

[33] New tools can be suggested for inclusion into the Switchboard by anybody, but they will be evaluated by the Switchboard maintainers before inclusion to ensure that only stable services are included.

[34] Programming Historian website: https://programminghistorian.org/en/; [20.10.2021]

[35] Standardization Survival Kit: http://ssk.huma-num.fr; [30.08.2021]

[36] PARTHENOS project: http://www.parthenos-project.eu/; [30.08.2021]

steps do not only explain in detail which action to take, they usually also refer to further material like documentation or tutorials and they give recommendations on which specific standards to use.

An initial set of scenarios was compiled during the runtime of PARTHENOS. Afterwards the DARIAH working group "Guidelines and standards"[37] took over the maintenance of the SSK and further scenarios were gathered during workshops funded by DARIAH. It is expected that more scenarios will be integrated into the SSK.

Scenarios from the SSK are imported in the SSH Open Marketplace as workflows, steps from the SSK are imported as steps of these workflows, and the referenced material - coming from a Zotero library[38] - is imported either as publications, training materials, tools or services. Currently the SSH Open Marketplace holds 29 workflows, 264 steps, and 372 referenced material from the SSK. Also the list of standards from the SSK is used as a vocabulary in the SSH Open Marketplace.

TAPoR

TAPoR[39] is short for Text Analysis Portal for Research. It is the longest-standing discovery platform for tools in the Digital Humanities, holding around 1.500 individual entries. It was founded by Canadian scholars Geoffrey Rockwell and Stéfan Sinclair in 2002. Originally started as a collection of text-analysis software, it has widened its scope with the demise of DiRT (Directory for Research Tools) whose entries it inherited. So far, 1.370 tools from TAPoR are featured in the Marketplace.

2.2 Sources to be ingested by the end of the SSHOC project

This section provides an overview of all sources that are planned to be ingested into the Marketplace in the following months until the end of the SSHOC project. Each of the sources is described in at most a few paragraphs. It was decided to leave out the more technical details of the implementation, e.g. the mapping of metadata fields into the Marketplace data model.

CESSDA Training resources

The CESSDA Training Working Group, implementing one of the four strategical pillars of CESSDA, offers a wide variety of training in Research Data Management and data archiving to both researchers and data curators. The material produced in support of these trainings is available from CESSDA's training website and will be made available for direct harvesting into the SSHOC Open Marketplace by the end of the project.

[37] DARIAH-EU working group "Guidelines and standards": https://www.dariah.eu/activities/working-groups/guidelines-and-standards/; [30.08.2021]
[38] Zotero group "SSK_Parthenos": https://www.zotero.org/groups/427927/ssk-parthenos; [30.08.2021]
[39] TAPoR 3 website: http://tapor.ca/home [20.10.2021]

CLARIN Resource Families

The CLARIN Resource Families[40] are a number of curated collections of corpora and tools. They are manually put together by CLARIN with the aim to provide a user-friendly overview of existing resources within and without the CLARIN infrastructure. For a lot of the entries in the resource families there is also a related publication being listed, which can be picked up during ingestion into the Marketplace. The Resource Families are very actively curated and it is therefore planned to run updating ingests of this source regularly to keep up to date with all changes.

DARIAH-Campus

DARIAH-Campus is both a discovery framework and a hosting platform for DARIAH and DARIAH-affiliated training and education materials. Born out of the H2020-funded DESIR (DARIAH-ERIC Sustainability Refined) Project, DARIAH-Campus seeks to widen access to open, inclusive, high-quality learning materials that enhance creativity, skills, technology and knowledge in the digitally-enabled arts and humanities. DARIAH-Campus contains both resources hosted on the website and external links to pre-existing resources. DARIAH-Campus provides a way to render ephemeral training workshops permanent via hosting of videos. These entries are evaluated and curated by DARIAH ERIC staff. The data is uploaded and maintained through GitHub and in the summer of 2021 DARIAH-Campus received a new content management system (CMS) overlay that allows for easier additions by contributors. Once this update was complete, DARIAH-Campus content descriptions were ingested into the SSH Open Marketplace.

DARIAH contribution tool

DARIAH member states contribute to the DARIAH distributed infrastructure with a diverse range of resources and services[41], and declare these in-kind contributions via the DARIAH contribution tool[42]. The metadata schema of the tool follows a Reference Architecture developed within the Humanities at Scale project[43] and allows for the description of each contribution with core metadata fields, before continuing the description and evaluation of the contributions through a self-assessment, supported by additional metadata fields, and a review step that gives a final score to the contribution.

Given the diversity of content - activities and services can be registered as contributions - in the tool, and the difference in scope, purpose and architecture between the Contribution tool and the SSH Open Marketplace, filtering the contributions before their ingestion in the Marketplace is a necessary step.

[40] CLARIN Resource Families: https://www.clarin.eu/resource-families; [30.08.2021]

[41] See DARIAH contributions presentation on the website - https://www.dariah.eu/tools-services/contributions/; [20.07.2021] - and for more details about DARIAH services: Barbot, L., Roi, A., Scharnhorst, A., Durco, M., Fischer, F., Kalman, T., Moranville, Y., Parkola, T., Garnett, V., Edmond, J., Toth-Czifra, E. (2021). Towards a concise DARIAH service strategy: 2020 Reflections - White Paper. Report. https://doi.org/10.5281/ZENODO.4621287 [10.12.2021]

[42] DARIAH contribution tool: https://contrib.dariah.eu/; [20.07.2021]

[43] Humanities at Scale project: http://has.dariah.eu/; [20.07.2021]

The tool is hosted at DANS[44] and maintained by the DARIAH-CIO team[45], it exposes its data via an API[46]. It is considered a source of importance to highlight the DARIAH national assets in the EOSC context, as not all resources of interest registered via the tool are meeting the EOSC catalogue entry requirements, but are still of interest for national and/or EU research communities in Arts and Humanities.

SSHOC SERVICE CATALOGUE

The SSHOC service catalogue[47] is the result of all SSHOC Work Packages (WP), coordinated by WPs 1 and 2, to collect and consolidate the SSHOC services (or resources) offered. Based on the progressing implementation of the project, resources referenced in the catalogue are the most visible outputs of the SSHOC project. Building this catalogue serves multiple goals: it is a way to identify the Key Exploitable Results of the project and prepare the SSHOC Exploitation Plan; it is also a preliminary and necessary step to the EOSC onboarding for the services listed there[48]; it contributes to a better dissemination of the project outputs (via the publication of services' factsheets for example).

Currently hosted on the SSHOC website, a wordpress instance managed by TRUST-IT as part of the WP2 activities, one of the goals in ingesting the SSHOC service catalogue in the SSH Open Marketplace is actually to transfer this catalogue from the SSHOC website to the Marketplace. After an initial ingest the plan is to rely on the Marketplace portal to support the next curation steps, and for the final release of the Marketplace to replace the current presentation on the SSHOC website by a dynamic inclusion of the SSHOC content from the Marketplace.

SSHOC TRAINING MATERIALS

Not only SSHOC WP6 produces training material but also other SSHOC WPs are creating such material. These outcomes are communicated on the SSHOC project website, on social media channels, and as documents they are often deposited at the Zenodo community space[49]. But there are also training videos that are uploaded to the Youtube channel[50] of SSHOC. The training material produced in SSHOC should also be discoverable and available via the SSH Open Marketplace.

As there are different points of access to the SSHOC training material, this source is not easy to cover. There are also duplicates possible with the data of the SSH Training Discovery Toolkit. The approach is to harvest the communication channels of SSHOC to identify the produced training material and refer to the access points. Ideally, such an ingest is run at the end of the SSHOC project.

---

[44] DANS, https://dans.knaw.nl/en; [20.07.2021]
[45] DARIAH-CIO team: https://www.dariah.eu/about/organisation-and-governance/#cio-t; [20.07.2021]
[46] DARIAH contribution tool API documentation: https://dans-labs.github.io/dariah-contrib/Workings/API/; [20.07.2021]
[47] SSHOC service catalogue: https://www.sshopencloud.eu/service-catalogue; [21.07.2021]
[48] As part of the EOSC-Enhance project activities, one of CESSDA's missions, representing the SSHOC cluster project, is to coordinate the EOSC onboarding for the SSHOC services.
[49] Zenodo SSHOC community space: https://zenodo.org/communities/sshoc/ [20.10.2021]
[50] SSHOC Youtube channel: https://www.youtube.com/channel/UCw-mY8v84yeHW2z4KG3ZLtA [20.10.2021]

SSH Conversion Hub

The SSH Conversion Hub is an outcome of SSHOC WP3 (See *D3.6 Report on SSHOC format interoperability solution services*[51]). It allows users to search for tools that convert from one (meta-)data/file format to another one, e.g. from CSV (comma-separated values) to TEI (Text Encoding Initiative). There are two types of items available: tools that offer conversions from one format to another; for those a record states among others information about the input and the output formats. On the other hand there are recipes that describe more complex conversion chains, e.g. involving several steps using more than one tool or manual intervention. The SSH Conversion Hub allows users to search for input or output formats and other service metadata and return both tools and recipes that meet the search criteria. A faceted result gives the possibility to further limit the search results.

The SSH Conversion Hub is currently being finalized. The data collection will be ingested into the SSH Open Marketplace when the SSH Conversion Hub becomes public. As the Conversion Hub data model was created in coordination with WP7 there is already a stable mapping in place. Also some of the created vocabularies of the SSH Conversion Hub are used for the SSH Open Marketplace, e.g. the "invocation type"-vocabulary. It is intended to import the conversion tools as tools/services in the SSH Open Marketplace and the recipes as workflows and steps.

SSH Training Discovery Toolkit

The SSH Training Discovery Toolkit[52] is an outcome of SSHOC WP6. It acts as an overview on relevant sources that hold (digital) material for trainers. For such sources, selected items are described in more detail than is useful for the SSH community and give a hint about what training material to expect from the source. Examples of sources are the CLARIN Legal Information Platform[53] holding items on licensing and data protection, DelftX[54] offering free online courses from Delft University of Technology, and the EOSC-Synergy Training Platform[55] with different tools, services and online training material to use EOSC.

The SSH Training Discovery Toolkit is still in development and further sources and items are to be included by SSHOC WP6. The data collection will be ingested into the SSH Open Marketplace, especially since the collected items do have high relevancy. Some of the sources may also be directly harvested by the SSH Open Marketplace to get all of the items from there. There is already an elaborated collaboration between WP7 and WP6 regarding the data model of the SSH Training Discovery Toolkit that will allow an easy mapping to the SSH Open Marketplace.

---

[51] Mari Kleemola, Katja Moilanen, Daan Broeder, Matej Ďurčo, Klaus Illmayer, Maurizio Sanesi, Emiliano Degl'Innocenti, Hervé L'Hours, Benjamin Mathers, Johan Fihn Marberg, Eleni Tsoulouha, Athina Kritsotaki, & Cesare Concordia. (2021). D3.6 Report on SSHOC format interoperability solution services, including new software. https://doi.org/10.5281/zenodo.5561604 [10.12.2021]

[52] SSH Training Discovery Toolkit: https://training-toolkit.sshoc.eu/ [10.09.2021]

[53] CLARIN Legal Information Platform: https://www.clarin.eu/content/legal-information-platform [10.09.2021]

[54] DelftX website: https://www.edx.org/school/delftx [10.09.2021]

[55] EOSC-Synergy learn platform: https://learn.eosc-synergy.eu/ [10.09.2021]

2.3 Discarded sources

This section lists all other sources that were initially considered for inclusion into the Marketplace (the so-called longlist). For each of them a short description and a clarification of the reasons behind this decision are provided.

Awesome Digital History

There are a lot of manually crafted lists on GitHub/GitLab that give an overview on training material, tools and services on a dedicated topic. Some of these lists refer to digital research methods useful for a discipline. One of them is the "awesome digital history"[56] list. It contains in a structured manner sources like archives and it also refers to learning resources to apply digital history methods on these sources. It also refers to further awesome-lists like the "Tools for Academic Research"[57] that are organised in a quite similar way based on GitHub/GitLab and usually using the markdown format, to describe the listed resources. Even though these awesome lists are in principle a valuable source for the SSH Open Marketplace, some problems were discovered with the ingestion pipelines. First, there is no agreed metadata standard for awesome lists, thus it requires a dedicated mapping definition for every such list. There is also a lack of metadata, usually only giving a short description, a title and the link to such a tool. Often the categorisation is organised by headers in the text and not in a metadata format. Therefore - secondly - it would have been needed to apply either some NLP method to gather good metadata quality or - like for the "awesome digital history" - would not have enough contextual data on the tools on the list (even not the information, if it is a tool or a training resource). Thirdly, some of these lists are outdated - the last update for "awesome digital history" was two years ago - therefore it is to be expected that extensive curation would be necessary after ingestion. Finally, even though awesome lists are interesting sources for the SSH Open Marketplace, T7.3 members thought that the ingestion of TAPoR already covered most of the tools and services from such lists.

B2FIND

EUDAT's metadata catalogue, B2FIND[58], aggregates information from research data collections from EUDAT data centres and other repositories. For the SSHOC community, its main relevance is the metadata harvested from the GESIS Data Archive and from some CLARIN centres. However, all of these records are also accessible through other portals, the CESSDA Data Catalogue (for GESIS) and the Virtual Language Observatory (for the CLARIN centres). Since these community-specific metadata catalogues will be included directly into the SSH Open Marketplace, it was decided to leave out B2FIND as an information source.

[56] Awesome digital History: https://maehr.github.io/awesome-digital-history/ [10.09.2021]
[57] Tools for academic research website: https://tools.kausalflow.com/tools/ [10.09.2021]
[58] B2FIND website: https://eudat.eu/services/b2find [10.09.2021]

DARIAH-DE Collection Registry

The DARIAH-DE Collection Registry "serves as a catalog of collections which occurred within the scope of research projects or serves as a basis for them; links data whose data models and the description of a collection for technical reuse by services such as search or analysis tools [and] serves to manage collection descriptions"[59]. It comprises around 200 collections that were identified as of interest for the SSH Open Marketplace, but was discarded as a source after some discussions that also helped to shape the perimeter of the Marketplace. First of all, the question raised was whether to add the Collection Registry as one entry or to "harvest" the collections and/or the repositories it references. Because of the quality of the metadata (often too little or only available in German) and the focus of T7.3 on contextualisation and links between items - esp. for the datasets and publications content types in the Marketplace - it was decided to not keep the DARIAH-DE Collection Registry in the list of sources to ingest. An additional reason for this decision was that the DARIAH-DE collections will also be soon included in the CLARIN VLO. Nevertheless, because of the integration of the Collection Registry in the wider "DARIAH-DE research data federation infrastructure"[60] registered as a DARIAH-EU in-kind contribution, this resource will be present in the Marketplace, albeit as a single entry.

ICPSR

The Inter-university Consortium for Political and Social Research (ICPSR) website[61] includes a "Teaching and learning" section with a set of resources for students and instructors eager to study or teach data analysis in social sciences. Useful resources are listed in dedicated pages for students[62] and teachers[63], and ~50 data-driven learning guides are selected and presented there[64]. Identified at the beginning of the SSHOC project as a useful training resource for social scientists, this source has been left out because of the changes introduced in the ICPSR website (URLs identified at first became obsolete while prioritising sources for ingestion). In the meantime, ICPSR materials have been included in the SSHOC Training Discovery toolkit[65] and will therefore be indexed in the SSH Open Marketplace via the Training Discovery Toolkit. Thus, the ICPSR website as such will not become a source for the Marketplace.

---

[59] DARIAH-DE Collection Registry description on DARIAH-DE website: https://de.dariah.eu/en/web/guest/collection-registry [21.10.2021]

[60] DARIAH-DE research data federation infrastructure description on DARIAH-DE website: https://de.dariah.eu/en/data-federation-architecturehttps://de.dariah.eu/en/data-federation-architecture [21.10.2021]

[61] ICPSR website: https://www.icpsr.umich.edu; [02.08.2021]

[62] ICPSR students resources: https://www.icpsr.umich.edu/web/pages/instructors/student-resources.html; [02.08.2021]

[63] ICPSR teachers resources: https://www.icpsr.umich.edu/web/pages/instructors/teacher-resources.html ; [02.08.2021]

[64] ICPSR Data-Driven Learning Guides: https://www.icpsr.umich.edu/web/instructors/biblio/resources; [02.08.2021]

[65] ICPSR Teaching and Learning entry in the SSHOC Training Discovery Toolkit: https://training-toolkit.sshopencloud.eu/source/327; [02.08.2021]

METHODICA

The Methodi.ca website[66] presents research methods and techniques and is the "TAPoR companion"[67]: while TAPoR collects tools, Methodi.ca explains how to use these tools based on concrete research use cases, also described as recipes. Because of this approach, the resources referenced in Methodi.ca have first been considered for inclusion as workflows in the SSH Open Marketplace. Nevertheless, considering that the website is not maintained anymore - a long list of error messages appears on each page - and that some of the topics covered by the recipes and tutorials are also covered by the Programming Historian lessons, the decision was taken to discard Methodi.ca from the data sources of the Marketplace.

TERESAH

Tools E-Registry for E-Social science, Arts and Humanities (TERESAH)[68] can be seen as the SSH Open Marketplace predecessor. This tool registry, built as part of the Data Service Infrastructure for the Social Sciences and Humanities (DASISH) project[69] and further developed by the Humanities at Scale (HaS) project[70], was mainly populated by the Digital Research Tools (DiRT) Directory[71] tools. Although TERESAH played a role in the very definition of the SSH Open Marketplace[72], it was not used as a data source, mainly because the records included there would have been duplicates of the ones already ingested in the Marketplace from TAPoR. The team considered ingesting TERESAH data to test the deduplication and merging mechanisms in the Marketplace, but implementation of these functionalities took longer than expected and this option was discarded.

2.4 Potential additional (re)sources

This section includes those sources that do contain interesting information that would be nice to have in the Marketplace, but for some reason an automatic ingestion of them is not possible (or in the case of OpenAIRE very complicated). Depending on the time left after ingesting all sources described in section 2.2, they could be added manually to the Marketplace once the curation user interface is fully functional.

---

[66] Methodi.ca: https://methodi.ca/; [20.07.2021]

[67] Methodi.ca presentation: https://methodi.ca/content/about; [21.07.2021]

[68] TERESAH website: http://teresah.dariah.eu/; [21.07.2021]

[69] The DASISH project is an FP7 project that brought together, between 2012 and 2014, the five ESFRI research infrastructure initiatives in the social sciences and humanities (SSH): CESSDA, CLARIN, DARIAH, ESS and SHARE. See https://www.cessda.eu/About/Projects/Past-projects/DASISH; [21.07.2021]

[70] Humanities at Scale project: http://has.dariah.eu/; [21.07.2021]

[71] The DiRT directory, developed as part of the US Bamboo project, is no longer maintained but (meta)data have been ingested into TAPoR before the directory was shut down.

[72] See Claudia Engelhardt, Claudio Leone, Yoann Moranville. Distributed Metadata Schema and Demonstrator for Open Humanities Methods. [Research Report] Göttingen State and University Library; DARIAH. 2017. ⟨hal-01637051⟩ [21.07.2021]

DH Toychest

The DH Toychest website[73] is one of the most well-known and long-standing DH resource registries curated by Alan Liu[74], University of California, Santa Barbara. It includes a variety of resources structured with the following categories: Guides to Digital Humanities; Tutorials; Tools; Examples; Data Collections & Datasets. Links are manually added in these different lists, but data are not structured enough to be automatically retrieved for ingestion in the SSH Open Marketplace. Nevertheless, because of the high level of curation some of the resources featured in DH Toychest are of high interest for the SSH Open Marketplace and should be considered for manual addition, once the curation components are implemented.

European Language Grid

The European Language Grid[75] (ELG) is a European project to develop and deploy a scalable cloud platform, providing, in an easy-to-integrate way, access to hundreds of commercial and non-commercial Language Technologies for all European languages, including running tools and services as well as data sets and resources. At the moment of writing the ELG platform contains some 5500 entries. About 335 of these are services, which can be called directly from the ELG web interface or via a standardized API. Accessing these services requires the registration of an ELG account.

Overall, given the amount of entries, the required authentication with the ELG platform, and the lack of a documented API[76] to harvest the metadata, the best option for integration within the SSH Open Marketplace seems to be as a single catalogue entry. In the longer term, there might be options[77] to connect the ELG services to the Switchboard. Even then, the choice for inclusion as a single catalogue entry would hold.

GLAM Workbench

The GLAM Workbench[78] is a collection of Jupyter notebooks that can be used to explore data coming from GLAM (Galleries, Libraries, Archives, and Museums) institutions, mainly from Australia and New Zealand cultural heritage collections. Specific tools, tutorials and datasets are gathered in the GLAM Workbench. Because of the integration of these resources in the Workbench, it is not an option to consider this website as a good candidate for massive data ingestion in the Marketplace. Nevertheless, content showcased there is of interest for researchers working on GLAM data so individual and manual additions (either of the workbench itself as a single record or of the tools and datasets used to build the workbench) in the Marketplace should be considered.

[73] DH Toychest: http://dhresourcesforprojectbuilding.pbworks.com/w/page/69244243/FrontPage; [21.07.2021]
[74] Alan Liu website: https://liu.english.ucsb.edu/ ; [21.07.2021]
[75] European Language Grid website: https://www.european-language-grid.eu/about/ [10.09.2021]
[76] European Language Grid documentation: https://european-language-grid.readthedocs.io [10.09.2021]
[77] github issue regarding options to connect the ELG and the Switchboard: https://github.com/clarin-eric/switchboard-tool-registry/issues/119 [10.09.2021]
[78] GLAM Workbench: https://glam-workbench.net/; [21.07.2021]

Innovations in Scholarly Communication

The Innovations in Scholarly Communication project website[79] presents the results of a global survey on individual research workflows and tool usage conducted in 2015-2016. 20,663 answers were gathered and helped building a series of analysis - from a shared database[80] to typical workflow examples[81] - of research activities. This work gives a very interesting overview of tool usage but is not field specific (even though a field specific variable can be used to filter relevant answers for social sciences and humanities). No further investigation has been conducted to reuse the tools collected in this survey, but it could be interesting to rely on the Open Science innovation levels used to categorise tools and workflows to create individual new entries in the SSH Open Marketplace and/or further describe the already ingested tools.

OpenMethods

The OpenMethods metablog[82] highlights Digital Humanities methods and tools by selecting and introducing already published content online. Developed as a DARIAH initiative, as part of the Humanities at Scale project[83] and in collaboration with OPERAS, this metablog collects highly curated content selected and introduced by the (OpenMethods) Editorial Team and categorised with the TaDIRAH taxonomy, in order to contribute to the "knowledge and critical discussion of digital methods and tools [that] is much needed to prove the value, chances and challenges of 'the humanities computing'."[84]. The 256 articles highlighted in OpenMethods[85] are good candidates to be added as publications in the SSH Open Marketplace and linked to the tools or workflows they relate to (or could also be manually added as `see-also` references by the Marketplace moderators). This source has not been prioritised for ingestion due to the lack of time and capacity available during the SSHOC project, but is a good candidate for continuous ingestion in the future.

Open Source Tools for Social Science Researchers

Spotted on Twitter by one of the SSHOC team members at the beginning of 2021, the Open Source Tools for Social Science Researchers collection[86] gathers ~140 tools and resources useful for social scientists. At the beginning of 2021, the SSH Open Marketplace was mainly gathering Humanities content, so this source has been seen as a valuable addition to Social Sciences resources. Because of the other sources prioritised, no further investigation about the granularity of resources provided (TAPoR is being listed next to Gephi) or the best way to ingest this collection into the Marketplace has

[79] Innovations in Scholarly Communication website: https://101innovations.wordpress.com/; [21.07.2021]
[80] Bianca Kramer & Jeroen Bosman, 400+ Tools and innovations in scholarly communication: https://docs.google.com/spreadsheets/d/1KUMSeq_Pzp4KveZ7pb5rddcssk1XBTiLHniD0d3nDqo [27.07.2021]
[81] Typical workflow examples: https://101innovations.wordpress.com/workflows/; [27.07.2021]
[82] OpenMethods: https://openmethods.dariah.eu/; [27.07.2021]
[83] Humanities at Scale project website: http://has.dariah.eu/; [27.07.2021]
[84] OpenMethods About pages: https://openmethods.dariah.eu/about/; [27.07.2021]
[85] Numbers as of 27.07.2021
[86] Open Source Tools for Social Science Researchers: https://open-source-social-science.github.io/; [27.07.2021]

been conducted. Considering the number of resources listed, moderators could be interested to cross-check the list with the Marketplace's content and manually add new entries or enrich existing ones thanks to this Open Source Tools for Social Science Researchers collection.

SERISS TOOLS

Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS), as one of the previous thematic cluster projects, provides 7 tools[87] which mainly focus on European Social Survey (ESS) use cases, developed by their consortium of research infrastructures in social sciences. These tools seem to be rich and mainly open-source and freely available, so it might be helpful for the SSHOC users to know about them and be able to use them. Here is a list of the tool names :

- Fieldwork Management System (FMS)
- Questionnaire Design and Documentation Tool (QDDT)
- Question Variable Database (QVDB)
- surveycodings.org
- Survey Management Portal/My EVS
- Translation Management Tool (TMT)
- Variable Harmonization Hub (VHH)

Some of the SERISS tools are further developed within SSHOC - such as the Translation Management Tool or surveycodings.org for example - and will therefore be included in the SSH Open Marketplace via the SSHOC service catalogue. For the others, and if they are still maintained, T7.3 recommends adding them manually in the Marketplace.

TOOLS, DATA OR SERVICES MENTIONED DURING T7.1 INTERVIEWS

As part of SSHOC *D7.1 System Specification - SSH Open Marketplace*[88], a series of 22 interviews with researchers and support staff members has been conducted to elaborate the user requirements. During these interviews, questions about digital research habits and tool usage were asked and some tools or other types of resources such as datasets or services were mentioned by interviewees. Based on these materials, a list could be set up and manual addition to the Marketplace could be done by the end of the project, in order to reflect real research habits. This method of getting close to research practices to understand what methods and tools are used is, in general, what guides the Marketplace data population, and as described in deliverables *7.5 Marketplace Governance*[89] and *7.4 Marketplace Data population & curation* one of the rationales that will guide the Editorial team work.

[87] SERISS tools for cross-national research: https://seriss.eu/training/tools/ [10.09.2021]

[88] Laure Barbot, Yoan Moranville, Frank Fischer, Clara Petitfils, Matej Ďurčo, Klaus Illmayer, ... Sotiris Karampatakis. (2019). SSHOC D7.1 System Specification - SSH Open Marketplace (Version 1.0). Zenodo: https://doi.org/10.5281/zenodo.3547648; [27.07.2021]

[89] Clara Petitfils, Suzanne Dumouchel, Nicolas Larrousse, Edward J. Gray, Laure Barbot, Arnaud Roi, Matej Ďurčo, Klaus Illmayer, Stefan Buddenbohm, & Tomasz Parkola. (2021). D7.5 Marketplace - Governance. https://doi.org/10.5281/zenodo.5608487 [10.12.2021]

OpenAIRE

OpenAire EXPLORE[90] is a catalogue allowing to explore the OpenAIRE Research Graph[91], one of the largest open scholarly record collections worldwide, aggregating around 450 million metadata records about publications, datasets, as well as software and other research products, putting them in relation to each other. It is also considered to be the primary catalogue for research output in EOSC, complementing the EOSC marketplace with its focus on services. Thus, given the all-encompassing scope (no discipline restrictions), in theory, OpenAIRE Research Graph should contain most of the information relevant for SSHOC Marketplace. Therefore it was considered as one of the main candidates for ingestion.

The evaluation and first attempts to harness this rich trove exposed a number complications and deficiencies:

- Quantity: with 450 million records from all kinds of domains and disciplines the dataset is way too general to be useful for the specific needs of SSH Open Marketplace.
  Even when restricting the search to "Social Sciences and Humanities" Community, still more than 200,000 records are returned, which is well beyond any means for manual curation
- Quality: Most of the records have only very minimal metadata (authors, description/abstract, links to the resources and keywords or subject headings). Fields which could be potentially based on vocabularies and thus suited for facetting feature too many different terms, which make the vocabularies of the Marketplace unusable.
- Focus on Publications: The by far prevalent type of research outcomes are publications and other documents. Even though these are also relevant for the Marketplace, the strategy is to restrict to those publications that can be linked to other types of resources, primarily tools.
- Relations: One hope for additional information from OpenAIRE were the relations among the entities, however OpenAIRE seems to concentrate on relations between research outputs and their creators and contributors, or funders (i.e. actors in various roles). Anecdotal evidence suggests that there are no relations between tools and e.g. the publications, they are mentioned in.

In an attempt to extract from the OpenAIRE dataset a subset relevant for the purposes of the SSH Open Marketplace, T7.3 members looked for mentions of tools (already known in the marketplace). Initial evaluation based on 1,369 tools suggests that for 62% of them some information is contained in the OpenAIRE dataset. However, given a simple string-based search, many of these seem to be false positives, e.g. "Archive-it" returns over 30,000 results, most of them clearly just using the common words "archive" and "it". Thus here again a thorough manual evaluation and curation would be needed, as a base for iterative refinement of the automated search procedures.

[90] OpenAire EXPLORE platform: https://explore.openaire.eu/ [20.10.2021]
[91] OpenAire Research Graph: https://graph.openaire.eu [20.10.2021]

Thus in summary, as rich as this resource seems to be on the first glance, it is too big and too complex to be easily digested and used for the specific purposes of SSH Open Marketplace. However, given its pivotal role in the EOSC ecosystem, it cannot be dismissed easily and further efforts to establish connection between OpenAIRE and SSH Open Marketplace are needed.

3. Outlook and future work

Even after the final public release of the Marketplace at the end of 2021 there is still a lot of work that can be done to improve the Marketplace. During the remaining months of the SSHOC project, the team will focus on adding the sources detailed in section 2.2. There is a detailed timeline for the ingestion of these sources, but this might evolve during the coming months, especially because the DACE ingestion pipeline is still being developed and it is likely that it will be discovered that it needs to be adapted to be able to handle all the different types of sources. Additionally some of the sources that still need to be ingested contain only a small number of items and it might be decided to forego the automatic ingestion for those and add them manually instead.

Already now it is one of the biggest outcomes of this task within WP7 that the sources that were targeted for ingestion are very heterogeneous regarding the number of items per source, the details provided for each item, the possibility to harvest the source (e.g. does it provide an API?) and the amount of curation being done within the source (e.g. are there a lot of dead links or features missing from a description, because the item record was not updated after a new tool version was released). Depending on all of these factors the amount of work needed to ingest the source varied accordingly.

After the project has ended, there will still remain a number of open issues, though. First and foremost, as can be seen in section 2.3.2, there are quite a number of interesting sources that could not be ingested due to lack of time and some of them would be very valuable additions to the Marketplace. Furthermore, even though a long list of possible sources has been collected, it is very likely that there are other interesting sources that T7.3 members didn't know about and which would be interesting to add. The Marketplace will continue to be open to further contributions. Apart from manual additions of items through the editorial interface (see *D7.4 Marketplace Data Population and Curation*), it might also be possible to ingest larger sources using the existing ingestion pipelines.

For this, potential sources should meet some minimum requirements as explained in further detail in the introduction. In short, potential resources will be evaluated regarding their **quality** (they should be useful additions), their **uniqueness** (they should not consist of items mostly already in the Marketplace) their **technical interface** (they should have metadata that can be relatively easily mapped to the Marketplace data model and should be harvestable via an API), how much they will enhance the **representativity** of the various SSH domains within the Marketplace, and how **useful** they will be to the users of the Marketplace. Additionally, it will be favourable if a new potential source will bring with it possible **contextualization**, i.e. relations to other items within itself or the Marketplace.

Indeed, keeping in mind the sustainability scenarios detailed in *D7.5 Sustainability and Governance*, it is vital that future ingestions are planned to maximize the human and technical resources available, in

order to ensure the most effective future data population for the SSH Open Marketplace. Chiefly, perhaps the most important criteria for evaluating future ingestion sources is their completeness, and level of maintenance. Any future ingestion should be done with a repository that has one, clean data model, is up-to-date and is well-maintained. Trying to resurrect data from obsolete and abandoned repositories, however noteworthy that data may be, is a significant challenge.

Beyond these factors, there are other considerations born out of the experience with ingestion during the lifetime of the SSHOC project. For instance, appropriate time must be dedicated to the complicated task of mapping the properties of a potential source with that of the SSH Open Marketplace, and special attention should be paid to the vocabularies that have been implemented in the Marketplace data model. Concurrently, it is vital to respect the mandatory fields of the SSH Open Marketplace data model. Any customization of the data ingest should happen at the level of mapping, and not during curation. Technically, sources are very unlikely to be included if it is necessary to use different and multiple API calls for a single ingestion.

Such a setup makes the task very difficult if not impossible based on the ingestion pipeline used. The same is true for sources that have hidden data not available via the API or other technical limitations that make ingesting them via the existing ingestion pipelines extraordinarily difficult. Such sources should be avoided. In the end the decision whether a new source will be ingested into the Marketplace will be made by the Marketplace Editorial Board[92] on a case by case basis using the factors outlined in this section as a basis.

The Marketplace is accessible through an API itself. This makes it easy to use programmatically and thus could open up possible further use cases which weren't foreseen by the SSHOC project yet. In particular, it could be used for purposes within the EOSC currently not known, e.g. a feedback loop of enriched metadata towards the EOSC Portal Marketplace.

There are also some open questions that need further investigation and discussion. First, as seen with some sources, most notably the Programming Historian, multilinguality is an issue that needs to be looked at. This is twofold, on the one hand source items can be in languages other than English, be it the whole item, e.g. for a workflow, the UI, e.g. for a software tool or just the description or documentation. Some of these items might be valuable for users of the Marketplace, but they have to be flagged in a certain way, so that it is possible to search for material only in specific languages. The final release of the Marketplace will have a language facet that will make it possible to filter items by language. On the other hand, it should be investigated if the Marketplace UI itself can be made available in languages other than English. This would in theory increase the potential user base and help users that do not have a good grasp of English, but for this to have a real impact there would need to be a larger number of non-English language entries in the Marketplace as well.

[92] See SSHOC D7.5 Marketplace Governance for more details on the Editorial Board.

Another open issue is how to deal with general purpose services or tools. There are a number of tools and services that can be useful for any discipline and not just the SSH (e.g. statistics software or Jupyter Notebooks[93]). The Marketplace is meant to serve specifically the SSH community and it needs to be discussed whether such general purpose tools should be included or not. See also the Editorial Guidelines, included as annex of *D7.4 Data Population and Curation* for a longer discussion of this point.

[93] EGI Notebooks on the EOSC Portal Marketplace: https://marketplace.eosc-portal.eu/services/egi-notebooks [21.10.2021]

4. References

Edward Gray, Nicolas Larrousse, Clara Petitfils, Laure Barbot, Frank Fischer, Matej Ďurčo, Klaus Illmayer, Cesare Condordia, Alexander Konig, Dieter Van Uytvanck, & Stefan Buddenbohm. (2021). D7.4 Marketplace – Data population & curation (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5783358

Laure Barbot, Yoan Moranville, Frank Fischer, Clara Petitfils, Matej Ďurčo, Klaus Illmayer, ... Sotiris Karampatakis. (2019). SSHOC D7.1 System Specification - SSH Open Marketplace (Version 1.0). Zenodo: https://doi.org/10.5281/zenodo.3547648

Laure Barbot, Arnaud Roi, Andrea Scharnhorst, Matej Durco, Frank Fischer, Tibor Kalman, Yoann Moranville, Tomasz Parkola, Vicky Garnett, Jennifer Edmond & Erzsebet Toth-Czifra. (2021). Towards a concise DARIAH service strategy: 2020 Reflections - White Paper. Zenodo. https://doi.org/10.5281/zenodo.4621287

Laure Barbot, Frank Fischer, Klaus Illmayer, Matej Ďurčo, Alexander König, Dieter Van Uytvanck, & Nicolas Larrousse. (2020). MS.43 Marketplace - beta release (1.0). Zenodo. https://doi.org/10.5281/zenodo.4785194

Daan Broeder, Willem Elbers, Stefan Buddenbohm, Wolfgang Schmidle, Emanuel Dima, Matej Durco, Cesare Concordia, Maurizio Sanesi, & Emiliano Degl'Innocenti. (2021). D3.8 Implementation report and available SSHOC Switchboard and VCR services (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5608542

Matej Ďurčo, Laure Barbot, Klaus Illmayer, Sotiris Karampatakis, Frank Fischer, Yoann Moranville, Joshua Tetteh Ocansey, Stefan Probst, Michał Kozak, Stefan Buddenbohm, & Seung-Bin Yim. (2021). D7.2 Marketplace – Implementation (v1.0). Zenodo. https://doi.org/10.5281/zenodo.5749465

Claudia Engelhardt, Claudio Leone, Yoann Moranville. Distributed Metadata Schema and Demonstrator for Open Humanities Methods. [Research Report] Göttingen State and University Library; DARIAH. 2017. https://hal.archives-ouvertes.fr/hal-01637051

Mari Kleemola, Katja Moilanen, Daan Broeder, Matej Ďurčo, Klaus Illmayer, Maurizio Sanesi, Emiliano Degl'Innocenti, Hervé L'Hours, Benjamin Mathers, Johan Fihn Marberg, Eleni Tsoulouha, Athina Kritsotaki, & Cesare Concordia. (2021). D3.6 Report on SSHOC format interoperability solution services, including new software. https://doi.org/10.5281/zenodo.5561604

Clara Petitfils, Suzanne Dumouchel, Nicolas Larrousse, Edward J. Gray, Laure Barbot, Arnaud Roi, Matej Ďurčo, Klaus Illmayer, Stefan Buddenbohm, & Tomasz Parkola. (2021). D7.5 Marketplace - Governance. https://doi.org/10.5281/zenodo.5608487

List of Figures & Tables