The SSH Open Marketplace and CLARIN

Alexander König
CLARIN ERIC
alex@clarin.eu

Laure Barbot
DARIAH-EU
laure.barbot@dariah.eu

Cristina Grisot
CLARIN-CH
DARIAH-CH
cristina.grisot@uzh.ch

Michael Kurzmeier
Austrian Centre for Digital Humanities
michael.kurzmeier@oeaw.ac.at

Edward J. Gray
DARIAH-EU
IR* Huma-Num
edward.gray@dariah.eu

Abstract

This paper showcases the SSH Open Marketplace, which is a discovery portal which pools and contextualises resources for Social Sciences and Humanities research communities, and its tight connections with the CLARIN infrastructure. The proposal presents how the SSH Open Marketplace can provide insights into the use of tools, methods and standards in the Social Sciences and Humanities communities in general, and for the CLARIN community in particular. The paper also describes how the SSH Open Marketplace can increase serendipity in the discovery of new methods and standards, by interlinking the resources and describing workflows. As contextualisation is provided between the items of the catalogue, it is easy to understand and assess the usefulness of a resource.

1 Introduction

In the context of Open Science, infrastructures, catalogues and discovery portals play a crucial role for enabling open research data and for increasing the degree of FAIRness (findability, accessibility, interoperability and reusability) of research data. The Social Sciences and Humanities Open Marketplace (SSH Open Marketplace) - marketplace.sshopencloud.eu - is a discovery portal which pools and contextualises resources for Social Sciences and Humanities research communities: **tools, services, training materials, datasets, publications and workflows**. The SSH Open Marketplace showcases solutions and research practices for every step of the research data life cycle. In doing so, it facilitates discoverability and findability of research services and products that are essential to enable sharing and re-use of workflows and methodologies.

The SSH Open Marketplace, conceptualized and implemented during the Social Sciences and Humanities Open Cloud (SSHOC) project [1], is one of the pieces of the bigger puzzle called the European Open Science Cloud (EOSC)[2]. The vision for EOSC is to create an environment for hosting and processing research data to support EU science, which provides seamless, Europe-wide access to research data and tools across scientific or thematic disciplines and geographical borders. The SSHOC project, along with the other four thematic clusters, namely ESCAPE (astronomy and particle physics), ENVRI (environmental sciences), panosc (materials, health, energy, physics) and EOSC-Life (life sciences)[3], supported the integration and consolidation of thematic e-infrastructure platforms in preparation for connecting them to the EOSC.

In this vein, the overall objective of the SSHOC project was to realise the Social Sciences and Humanities component of EOSC. As a domain-oriented discovery portal and the aggregator of the SSHOC project, the SSH Open Marketplace, contributes directly to the EOSC, supplementing existing services

[1] See the SSHOC project description: https://cordis.europa.eu/project/id/823782

[2] See EOSC description on the European Commission Directorate-General for Research and Innovation website: https://research-and-innovation.ec.europa.eu/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en

[3] See EOSC portal about the five ESFRI cluster projects in the EOSC panorama: https://eosc-portal.eu/news-and-events/news/five-new-esfri-cluster-projects-eosc-panorama

such as the EOSC Catalogue and Marketplace, and facilitating the fluid exchange of tools, services, data, and knowledge. As a continuation of the SSHOC project and to sustain its outputs, 5 ESFRI Landmarks, CESSDA, CLARIN, DARIAH, ESS and SHARE, have signed a Memorandum of Understanding for the establishment of the **SSH Open Cluster**, and were later joined by 14 other national or European institutions and/or research infrastructures. All ESFRI projects and landmarks from the Social and Cultural Innovation domains [4] are currently members of the SSH Open Cluster. This cluster acts as an umbrella for the SSH Open Marketplace organisation and activities. More generally, the collaboration between the SSH Open Marketplace stakeholders (funders, providers, moderators or contributors) ensures that these cataloguing and contextualising efforts are meaningful, notably because they are undertaken by and serve humanities researchers.

The SSH Open Marketplace is one of the 33 Key Exploitable Results of the SSHOC project, and CLARIN, DARIAH and CESSDA decided to ensure the sustainability of the service after the end of the project. They act as a Governing Board for the SSH Open Marketplace and define the Marketplace strategic policy with regards to scientific, technical and managerial matters. In that context, two institutions act as service providers on behalf of these ERICs: the Austrian Centre for Digital Humanities and Cultural Heritage (ACDH-CH)[5] of the Austrian Academy of Sciences providing hosting and maintenance for the service; and the Poznan Supercomputing and Networking Center (PSNC)[6], affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences, providing the data ingestion pipeline as well as maintenance for the service. The SSH Open Marketplace can also count on an Editorial Board, composed of 17 members[7], to ensure the day-to-day maintenance and (meta)data quality. Liaising with service providers and the end-users of the service (SSH researchers and support staff for researchers), the Editorial Board ensures the technical running of operation, the effectiveness of the curation process and the editorial policy's successful implementation.[8].

In sum, CLARIN has been heavily involved in the SSHOC project and is a founding partner for the continuation of the project in the form of the SSH Open Cluster. The SSH Open Marketplace is one of the key elements of research empowerment and discovery with which CLARIN is concerned. The special focus on contextualistion of the resources in the Marketplace can act as a complementary discovery tool to CLARIN's Virtual Language Observatory (VLO) - which includes a much larger number of items, but presents a lot less context - and the CLARIN Resource Families - which contain a much smaller number of items, but therefore can be even more extensively curated and contextualised.

2 Presentation of the SSH Open Marketplace

Initiated in the Digital Humanities (DH) context, inspired by the DiRT directory (Dombrowski, 2014), TAPoR[9] or the Standardization Survival Kit (Riondet & Romary, 2018) and aggregating their data for its initial population, the SSH Open Marketplace now acts as one of the thematic entry doors into EOSC. The SSH Open Marketplace was influenced by Dombrowski's ”directory paradox”, according to which DH tool registries should be community-led despite the organisational and infrastructural challenges it brings to such projects. (Dombrowski, 2021) This paradox is the foundation of one of the three Guiding principles which govern the SSH Open Marketplace, namely *Community* (see next section).

2.1 Guiding principles

While planning and building the SSH Open Marketplace three main pillars were identified, and these remain essential for its ongoing operation and future development. These pillars are:
Curation - The service thrives on a curation process that makes it easy to discover the most appropriate

[4] See the European Strategy Forum on Research Infrastructures website: https://roadmap2021.esfri.eu/projects-and-landmarks/

[5] see ACDH-CH website: https://www.oeaw.ac.at/acdh/acdh-ch-home

[6] see PSNC website: https://www.psnc.pl/

[7] see this page on the Marketplace website: https://marketplace.sshopencloud.eu/about/team

[8] For a detailed version of the sustainability plan, the report on Marketplace governance (Petitfils et al., 2021) can be consulted

[9] TAPoR 3: https://tapor.ca/. The Text Analysis Portal for Research is a project led by Geoffrey Rockwell and Milena Radzikowska.

and up-to-date results for each request, so that researchers can discover the best resources for the digital aspects of their work. The curation process relies on three components: automatic ingest and update of data sources; continuous curation of the information by the editorial team and – most important – contributions from users, the SSH research community.

Community – The content available in the SSH Open Marketplace and its contextualisation is the result of collaborative work that is characterised by a user-centric approach. Features that allow contributions are implemented to ensure that the portal mirrors real research practices.

Contextualisation – The portal puts all items into context: each solution suggested is linked to other related resources (e.g. a tutorial showing how to use a tool, a tool used in a workflow, a publication presenting research results produced using a given service). This contextualisation enhances the usefulness of the SSH Open Marketplace by showing how all these parts of the research process intertwine, and ensures users receive the maximum possible benefit from all its contents.

These three guiding principles describe the essence of the SSH Open Marketplace and contribute to increasing its usefulness for the target research communities and its sustainability in time.

2.2 Inclusion criteria

In order to guide users who wish to add resources to the SSH Open Marketplace, the following inclusion criteria and related guiding questions are enforced:

The relevance of the resource. The question to ask is: *will this resource be relevant to the SSH scientific community?* Thus, to be selected, any resource must fulfil at least two criteria: (1) scientific relevance and usefulness for SSH research and researchers and (2) pertinence to the digital methodologies used within the SSH landscape.

The technical status of the resource. The question to ask is: *is the resource current, supported, and ideally open?* The SSH Open Marketplace favours the uptake of Open Science workflows and open research practices. Software resources are preferably built upon open source solutions. Nonetheless, given that the SSH Open Marketplace seeks to mirror actual research practices, commercial or non-current resources are also referenced where these are relevant for the scientific community.

The degree of compliance of the resource with Open Science requirements. The question to ask is: *is the resource FAIR – Findable, Accessible, Interoperable and Re-usable - or contributing to the uptake of Open Science best practices?* The SSH Open Marketplace maximises the findability and re-use of data, and guides users towards tools, services or training materials that can help them in their FAIRification of workflows [10].

The uniqueness of the resource. The question to ask is: *is the resource already in the Marketplace?* If yes, there is no need to add it again, either as an individual item or with a source. Users are invited to enrich these existing items instead. However, when duplicates exist they can be identified as part of the curation activities (see Moderation and Curation section below). Currently, automatic checks of duplicates based on the name and the URL of the resources are regularly performed to merge the identified records. More parameters could be used in the future to reinforce the ”uniqueness of the resource” criterion.

Thanks to these inclusion criteria, the quality and the relevance of the resources added on the SSH Open Marketplace are guaranteed. This is an essential advantage for researchers who use the SSH Open Marketplace to discover resources which originate not only in their discipline but also from outside their own discipline. For example, a scholar who studies history and who uses digital methods to examine old documents, can easily discover on the SSH Open Marketplace the *Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools.* [11] This training material presents, in a Jupyter notebook format, a three-chapter tutorial - (1) XML and CMDI introduction, (2) Data selection and resource access, (3) NLP processing - allowing interested trainees to interact with the Europeana

[10] While the FAIR principles are very well suited to data, using them to evaluate software, services or other resource types does not make much sense. This is why these principles are mentioned here as criteria alongside other dimensions of Open Science, in order to encompass all resource types covered by the SSH Open Marketplace.

[11] Jupyter notebooks for Europeana newspaper text resource processing with CLARIN NLP tools. Version 1 Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/training-material/duVII1

newspaper collection using, for example, Named Entity Recognition tools from the CLARIN environment.

A lot of interesting resources are created to support SSH researchers in the digital aspects of their work. The multiplicity of resources in a fast-changing environment makes it sometimes difficult to keep track of the most relevant ones. The inclusion criteria are here to provide a generic framework for the items selected and published in the SSH Open Marketplace, and make sure that pertinent resources are populating the platform.

2.3 Item types

Taking into consideration user requirements and developments of the EOSC data model, 5 main content types have been identified to structure the SSH Open Marketplace resources (Barbot et al., 2021). These 5 item types are considered to be representative for the large array of digital resources that can be found on this discovery platform.

Tools and services which refer to services and products, such as software, applications, programs, websites, programming libraries and APIs. The trainable pipeline for tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files *UDPipe*[12] is an example of a tool provided by CLARIN.

Training materials are tutorials, lessons or didactic resources explaining how to perform an action or highlighting the potential learning outcomes gained from using that material. For example, the *CLARIN Hands-on Tutorial on Transcribing Interview Data* [13]focuses on the role of automatic speech recognition – what are the opportunities, what are the pitfalls and where can it be applied successfully.

Workflows are sequences of steps that one can perform on research data during their lifecycle. Workflows can be created by using diverse tools, resources and methods, and useful resources are connected to each step. For example, *Intertextuality phenomena in European drama history* [14]is a workflow composed of 4 steps useful for analysing the relationships between the characters in a drama based on monologue/dialogue.

Datasets are defined as an organised collection of data. They are generally associated with a unique body of work, typically covering one topic at a time and are treated as a single unit by a computer. The SSH Open Markeptlace indexes CLARIN Resource Families datasets, for example the *DK-CLARIN Reference Corpus of General Danish* [15]

Publications are defined as research results published in academic journals or non-peer-reviewed publication repositories such as Zenodo. The SSH Open Marketplace references only publications that can be connected to other resources (i.e. tools and services, training materials, workflows or datasets). For example, you can find a paper on *Using TEI, CMDI and ISOcat in CLARIN-DK*[16] or the *Dublin Core Metadata Schemas* [17] on the SSH Open Marketplace.

Despite the fact that these five content types are equally (re)presented in both the SSH Open Marketplace data model (Ďurčo et al., 2021) and the front-end of the service, metadata quality for some of the types and user interactions these last two years have led to a focus on tools and services, and how they can be contextualised thanks to the other content types, and as part of workflows. Training materials, datasets and publications are used and considered in relation to other items, while tools and services are more and more seen as the primary content type or ”first class citizen” of the SSH Open Marketplace. For example, when interacting with users during hands-on sessions to create or enrich items, one of the main questions is often ”which tool do you use to perform your research?”. This then leads to an elaboration

[12] UDPipe. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/tool-or-service/F7K42P

[13] SSHOC Webinar: CLARIN Hands-on Tutorial on Transcribing Interview Data. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/training-material/ITNpCC

[14] Intertextuality phenomena in European drama history. Retrieved Apr 27, 2023 from https://marketplace.sshopencloud.eu/workflow/DMJlzG

[15] DK-CLARIN Reference Corpus of General Danish. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/dataset/XR876U

[16] Dorte Haltrup Hansen, Lene Offersgaard, Sussi Olsen (2022): Using TEI, CMDI and ISOcat in CLARIN-DK. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/publication/4jQvZ5

[17] DCMI Schemas. Retrieved Sep 3, 2023 from https://marketplace.sshopencloud.eu/publication/6kYac0

on the context of use, giving shape to the creation or enrichment of the most relevant resources in the catalogue.

2.4 Moderation and Curation

With a population of approximately 7000 items, aggregated from more than 15 trusted sources, the SSH Open Marketplace relies on community curation - i.e. contributions from the research communities in SSH and from the Editorial Board - to ensure the catalogue entries remain up-to-date and useful for SSH researchers, the end-users of the portal.

Contributions from the research community can take the form of creation of new items or enrichment of existing ones. In both cases, contributors suggest changes that are then passed on to the moderators, i.e. the members of the Editorial Board, who accept or reject the suggestions. Rejection usually goes hand in hand with contacting the contributor and asking further questions or suggesting options to revisit the initial approach. Moderating is not reviewing, and this is why, especially in the case of workflow moderation, the checks performed are limited to editorial control rather than peer reviewing.

Furthermore, curation routines, mixing automatic and manual tasks, are set up to ensure and continuously improve (meta)data quality. Indeed, in order to gain an overview of the SSH Open Marketplace data and to perform some analysis to prioritise the curation tasks and improve the Marketplace data quality, a Python library and a set of Jupyter notebooks have been created[18]. The flexible scripts allow moderators and administrators to query the SSH Open Marketplace with advanced parameters and filters and, in some cases, to write back to the system to flag some items for curation in the editorial dashboard.

Close to the inclusion criteria, a set of quality criteria have been established - general entry requirements; non-redundancy; completeness of item description; verification of conformity and relevance; interlinking - to guide the improvement of the metadata quality[19]. Based on these criteria, quality metrics have been derived and are used as a basis for the checks performed via the notebooks. In practice, the curation tasks performed these last two years revolved around data monitoring, bug fixing and data enrichment. Particular attention has been paid to underlying elements such as the (controlled) vocabularies[20], actor curation[21] or relations between items. For instance, Editorial Board members have worked on the consolidation of the *keyword* vocabulary, an open vocabulary (or ’folksonomy’) in which users are allowed to add candidate concepts, and to which a wide variety of metadata from Marketplace sources have also been mapped, resulting in a 2000-concepts vocabulary. The curation exercise consists, in this case, in automatically mapping concepts belonging to other existing vocabularies in the Marketplace, or merging values when they are variations of the same concept. The work on the keyword vocabulary has also led to fruitful exchanges with other catalogues from the SSH domains as to how these services deal with ”topical vocabularies”. [22]

An important aspect to highlight here is that the manual curation work could not be done without the help and expertise of the members of the Editorial Board. As the SSH Open Marketplace covers a range of disciplines, the Editorial Board needs to mirror this diversity to be able to appropriately assess and make decisions on discipline-specific issues such as keywords, time periods or intended audience. This is why the approach of mixing automatic and manual curation is seen as a powerful one, that has proven its efficiency for a domain-oriented catalogue such as the SSH Open Marketplace in which rich metadata is essential to ensure findability of the resources.

[18] This library and the set of notebooks have been created by Cesare Concordia (CNR-ISTI) and are available at: https://github.com/SSHOC/marketplace-curation

[19] see SSH Open Marketplace moderator guidelines:https://marketplace.sshopencloud.eu/contribute/moderator-guidelines

[20] the SSH Open Marketplace currently counts 13 vocabularies, see:https://marketplace.sshopencloud.eu/contribute/metadata-guidelines

[21] In the Marketplace an Actor is the entity representing persons or institutions involved in the creation or maintenance of a resource. The SSH Open Marketplace counts around 7000 actors.

[22] see the TRIPLE project event *Use of vocabularies for metadata curation and quality assessment in Social Sciences and Humanities.* https://campus.dariah.eu/resource/events/use-of-vocabularies-for-metadata-curation-and-quality-assessment-in-social-sciences-and-humanities

3 The SSH Open Marketplace and CLARIN

3.1 CLARIN resources within the SSH Open Marketplace

The SSH Open Marketplace has been populated from a wide variety of sources, among which two come from the CLARIN world. The linguistic tools from the Language Resource Switchboard (LRS) (Zinn, 2018) and the tools, corpora and lexical resources collected in the CLARIN Resource Families (CRF) (Fišer et al., 2018). In both cases the original metadata has been mapped to the Marketplace Data Model (Ďurčo et al., 2021). As both the LRS and the CRF are very active, which means that items are constantly being added or updated (and in some cases also removed), the SSH Open Marketplace team has decided for a continuous ingest, i.e. to regularly re-harvest them to reflect changes at the source in the Marketplace.

Among the advantages of having the CLARIN resources listed in the SSH Open Marketplace is their increased discoverability by scholars from the SSH field who are not used to working with language data, who have not considered using language data in their research or who are not aware of the fact that CLARIN is an infrastructure which offers data and tools to support research that goes well beyond the linguistic domain. Indeed, language as cultural and social data is of interest for scholars from numerous other SSH disciplines. A historian, for instance, may use the SSH Open Marketplace discovery platform to find, in one search, language resources and tools (from CLARIN), digital resources from arts and humanities (from DARIAH) and social sciences resources (from CESSDA). Leveraging the SSH Open Cluster position in the EOSC to push these resources, useful beyond their initial target user group, to other audiences while maintaining the quality and richness of the accompanying metadata, even when processed through aggregation pipeline(s) is a challenge *per se* on which we elaborate in the last section of this paper.

Another important type of resource present on the SSH Open Marketplace is that of training materials. To increase their discoverability, such materials are being added to the SSH Open Marketplace, either manually or via the SSH Training Discovery Toolkit[23], which is also a source for the Marketplace. Giving space to and increasing the visibility of training and education is essential when it comes to increasing the access to open educational resources on various topics, and this focus is aligned to lines of action already present in the underlying ERICs, such as CLARIN and DARIAH. In what concerns CLARIN, it has recently strengthened its focus on training, especially in the context of the UPSKILLS project[24], which resulted in the creation of an important number of training resources, the creation of the CLARIN Learning Hub[25], and the creation, in collaboration with DARIAH, of the DH Course Registry[26].

3.2 Metadata and Interoperability

Metadata in the SSH Open Marketplace is subject to a process of mediation between the data available from the ingested source and the data requirements of the Marketplace users. The goal of this process ideally is to represent all important metadata taken from the ingested source while still maintaining a universal metadata structure in the Marketplace.

To illustrate, a tool from a source such as the CLARIN Language Resource Switchboard [27] can only be represented in the Marketplace if the input and output data formats can be recorded in the metadata describing the tool. While this is easy to achieve on a technical level, the universal approach on the Marketplace adds some additional steps to this process. For one, the chosen approach should be universally applicable to future ingests as well as manual curation of existing items. This likely means to create a controlled vocabulary from the data formats, so that different descriptions of the same format (for example, *XML file* and *TEI file* may be incorrectly used to describe the same format) can be unambiguously mapped to the canonical one. This new metadata then needs to be added as a metadata field in the back end and as searchable fields in the Marketplace front end. Additionally, the curation guidelines need

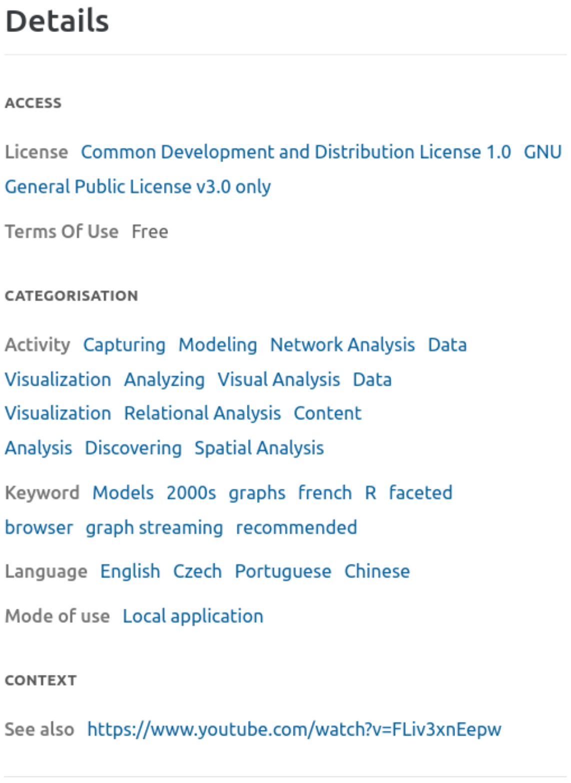
[23] see https://training-toolkit.sshopencloud.eu/entities?search=clarin

[24] see https://upskillsproject.eu/

[25] see https://www.clarin.eu/content/learning-hub

[26] see https://dhcr.clarin-dariah.eu/

[27] see https://switchboard.clarin.eu/



Details

ACCESS

License Common Development and Distribution License 1.0 GNU General Public License v3.0 only

Terms Of Use Free

CATEGORISATION

Activity Capturing Modeling Network Analysis Data Visualization Analyzing Visual Analysis Data Visualization Relational Analysis Content Analysis Discovering Spatial Analysis

Keyword Models 2000s graphs french R faceted browser graph streaming recommended

Language English Czech Portuguese Chinese

Mode of use Local application

CONTEXT

See also https://www.youtube.com/watch?v=FLiv3xnEepw

Figure 1: SSHOMP item detail with metadata

to reflect this new field and the user interface for manual data entry must highlight this important new attribute.

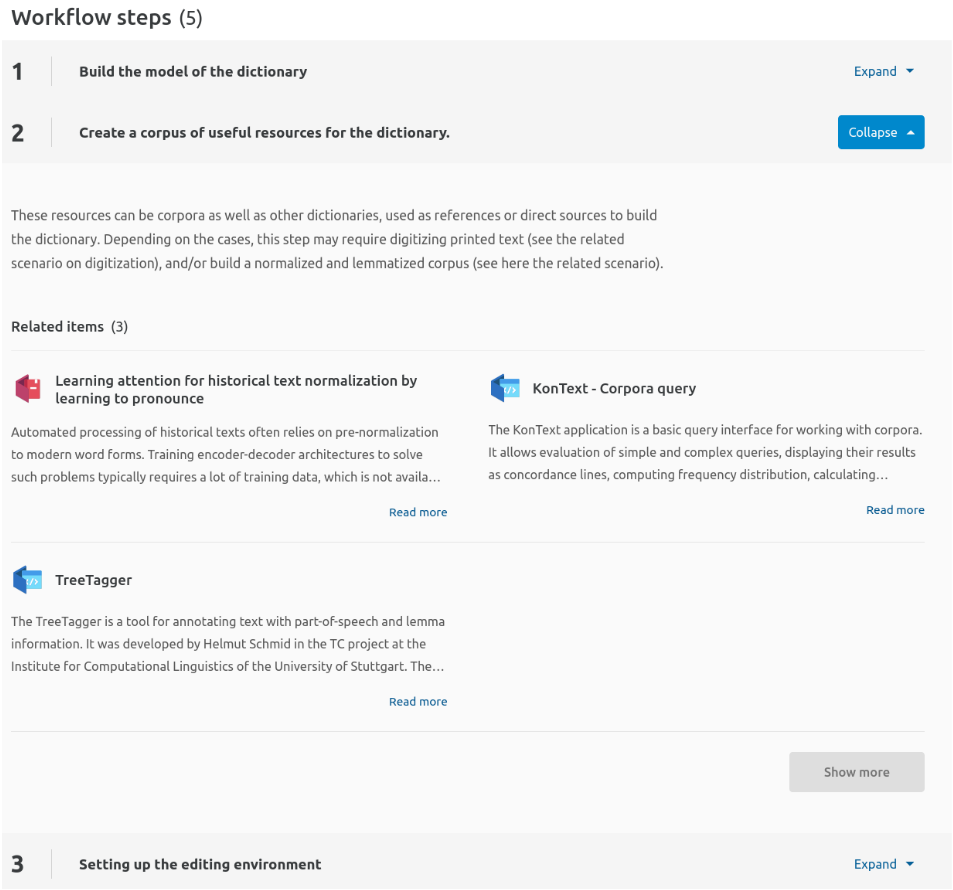
The described steps are to ensure that the new metadata field for data in- and output can be applied across the Marketplace. This in turn means employing resources to create the controlled vocabulary, changing the back and front end and adding additional steps to the item curation process. The example of the CLARIN Language Resource Switchboard serves to illustrate how the SSH Open Marketplace metadata structure can be adapted to different sources, however this process must always be undertaken with a view towards the generalist approach of the Marketplace.

Following the general process outlined in this example, the metadata schema of the SSH Open Marketplace can be adapted to accommodate new data sources as well as new types of interactions. In the example of the CLARIN Language Resource Switchboard, these new interactions are toolchains which enable multiple tools to be connected through shared data formats. In other cases, such as the SSHOC Conversion Hub [28], these interaction are conversions. Both require unambiguous metadata regarding the data input and output formats.

As can be seen in the above figure, metadata records in the SSH Open Marketplace are partially visible in the user interface, where they help to contextualize items. In the above example, all visible metadata fields except *Terms of Use* are clickable links through which the user can find other tools with the same attributes. This example further shows how important metadata records are for creating interoperability between items. Especially since the Marketplace combines different sources, universal metadata records create higher interoperability between items from different sources.

Being able to adapt the metadata structure of the Marketplace is essential to be able to represent ingested data from a variety of sources. The metadata structure also helps to express relations or common-

[28] see https://conversion-hub.sshopencloud.eu



Workflow steps (5)

1 Build the model of the dictionary

2 Create a corpus of useful resources for the dictionary.

These resources can be corpora as well as other dictionaries, used as references or direct sources to build the dictionary. Depending on the cases, this step may require digitizing printed text (see the related scenario on digitization), and/or build a normalized and lemmatized corpus (see here the related scenario).

Related items (3)

Learning attention for historical text normalization by learning to pronounce

Automated processing of historical texts often relies on pre-normalization to modern word forms. Training encoder-decoder architectures to solve such problems typically requires a lot of training data, which is not availa...

KonText - Corpora query

The KonText application is a basic query interface for working with corpora. It allows evaluation of simple and complex queries, displaying their results as concordance lines, computing frequency distribution, calculating...

TreeTagger

The TreeTagger is a tool for annotating text with part-of-speech and lemma information. It was developed by Helmut Schmid in the TC project at the Institute for Computational Linguistics of the University of Stuttgart. The...

3 Setting up the editing environment

Figure 2: Workflow detail

alities between items, such as the shared data formats, and thereby can help users find appropriate tools through an increase in searchable attributes. Decisions on new metadata fields are generally made with the involvement of the technical, curation and ingest teams of the SSH Open Marketplace (i.e. service providers and Editorial Board members).

3.3 Workflows

Workflows are, as described earlier, sequences of steps utilizing different SSH Open Marketplace items such as tools, training materials and datasets. Workflows have important functions in the SSH Open Marketplace, which will be described here insofar as they are in relation to CLARIN.

Connecting multiple items, workflows play an important role in contextualizing resources throughout the SSH Open Marketplace. Along all marketplace item types, workflows have by far the highest number of connected items, making them essential in contextualizing resources. Through this contextualisation work, previously unconnected items from different data ingests can be integrated into rich workflows.

Workflows allow researchers to present and share their methodology in a unique way. Through workflows, researchers can document every step of a methodology in accurate detail, thus providing a reproducible and transparent way of documenting research. Regarding research data management and FAIR principles, workflows can offer a compatible way to document and share research. Workflows are highly connected within the marketplace, and thus they allow for expert-led serendipity in the sense of discovering related resources that have been curated by the workflow author.

As the above Figure shows, workflows are composed of steps, which relate to different item types. Users can add as much detail as they like to workflows and workflow steps, making the workflow type ideal for documenting complex methodologies.

Because workflows combine practical application of individual tools with subject-matter expertise concerning the larger research project, they enable the SSH Open Marketplace to go beyond the representation of technical aspects about research tools and to provide the researcher with methodologies. This helps the SSH Open Marketplace to keep the focus on research and treat tools as utilities for said research workflows.

For instance, Marongiu et al., 2024 describe two multilingual workflows for semantic change research built on the SSH Open Marketplace and using CLARIN resources. The workflows proposed by Marongiu et al., 2024 aim to support research in lexical semantic change, i.e. the phenomenon by which words change their meaning over time. The workflows each consist of a series of steps required to detect words that have undergone semantic change as evidenced by a corpus and cover a range of user scenarios, including lexicology, historical research, and legal studies. The workflows are both research domain- and language-independent, and present the advantage of ”simplifying access to relevant language resources and tools scattered across different repositories and platforms” (Marongiu et al., 2024, section ”Reuse potential”). As put by the authors, the SSH Open Marketplace was chosen to be used as an environment for the creation of these workflows thanks to: (i) its possibility to create links to various resources for each step and even at the workflow level itself, provided that the resource in question is part of the platform, (ii) its broad scope across the whole SSH domains and its robust infrastructure, and (iii) its anchoring in three ERICs: CLARIN, DARIAH and CESSDA. The example of Marongiu et al., 2024 brings into light the great potential of the SSH Open Marketplace and its *workflow* type of item for increasing the findability and reusability of CLARIN resources.

At the time of writing, workflows in the SSH Open Marketplace are linear sequences of steps, and as such do not in all cases represent real-life research activities. A possible future extension of the Marketplace may therefore be the introduction of more flexible workflow types, including elements such as decision points, iterations and multiple end points. Workflows and their possible future extensions are discussed in more detail in a forthcoming publication. (see Barbot et al., 2024) These additions will increase the usability of the workflow item type in documenting existing research practices and will help to improve the usability of provided training resources.

3.4 Multilinguality of user interface and of records

Especially in the context of CLARIN resources, support for multilingual content in the SSH Open Marketplace is crucial. The SSH Open Marketplace currently features only an English user interface, with the majority of records presented in English. Operating as a monolingual discovery platform has its advantages and drawbacks. On the positive side, there’s enhanced discoverability of resources when users employ English keywords and the utilization of an English-controlled vocabulary. Conversely, the drawback lies in the necessity to translate names and descriptions of resources from non-English languages, recognizing that these resources may not be valuable to users unfamiliar with the resource’s language. Some of the vocabularies used in the SSH Open Marketplace, such as TaDiRAH, the Taxonomy of Digital Research Activities in the Humanities (Borek et al., 2021), are available in multiple languages and the Marketplace could also rely on them to bring more multilinguality to its interface.[29]

To evaluate the impact and potential expansion of the SSH Open Marketplace through multilinguality, it is crucial to comprehend the different levels of multilinguality in this context. Given that the marketplace serves as an aggregator and discovery platform, its content is metadata, not the resource itself, the resource content, documentation, and associated materials would still remain in its original language, most often English. Envisioning a scenario where users find translated metadata useful while the resource itself is solely available in English is challenging. For a more cohesive user experience, providing metadata in languages other than English becomes essential. However, as a resource aggregator, the SSH

[29] In that regard, the work done on the Triple Vocabulary (Triple Project Consortium, n.d.) based on a subset of the Library of Congress Subject Headings and with concepts translated in Greek, French, Polish, German, Italian, Portuguese, Spanish and Croatian is an interesting line of work to build on. See also the freely, openly available data a set of multilingual metadata concepts and an automatically extracted multilingual Data Stewardship terminology developed and delivered by Gamba et al., 2022

Open Marketplace currently cannot facilitate such translation work.

A solution would be to integrate an automatic translation system which would provide multilingual descriptions of the records. This would likely result in imperfect translations, which would need to be revised afterwards. This track was explored during the SSHOC project by Gamba et al., 2022, who tested the application of Natural Language Processing and Machine Translation approaches in view of providing resources and tools to foster multilingual access and discovery to SSH content across different languages. They tested state-of-art machine translation tools, such as Deep-L[30] or Google Translate[31] for the translation of metadata concepts and their definitions into Dutch, French, Greek, and Italian. Their exploratory works revealed both advantages and drawbacks to integrating automatic translation systems into the SSH Open Marketplace. On the one hand, the tested tools proved to be useful for the translation of metadata as the decrease of translation quality is minimal compared to the gain in terms of time and effort needed for traditional human translation. On the other hand, the results of the automatic translation must be revised and validated by humans who are experts in the domain and who have high language proficiency. In conclusion, current machine translation technologies do have limitations and cannot completely replace manual revision.

Furthermore, the Editorial Board encourages the inclusion of records in other languages, offering suggestions to enhance the search experience for non-English records. Specifically, it is recommended to use or at least add English names for entries to aid discoverability. If an English name doesn't exist or translating it doesn't make sense, including an English description in the title is recommended (e.g., Portal xx, Corpus xx). Resource descriptions can be in a language other than English, but including a short English description is advised for broader discoverability. Additionally, the use of English keywords is strongly encouraged for consistent discoverability, and specifying the language of the resource in the dedicated metadata field is highly recommended.

Finally, multilinguality takes on yet another meaning as it can refer to languages a tool or service is capable of processing. This broader perspective underscores the importance of accommodating diverse linguistic needs within the SSH Open Marketplace. While acknowledging the importance of multilinguality, the Editorial Board defers this aspect to future work for the SSH Open Marketplace.

3.5 Towards a better integration

First, investigating how a better connection between the CLARIN Resource Families and the SSH Open Marketplace could take shape is one of the main lines of work for the Editorial Board and the CLARIN team involved. At the moment, indeed, CRF records are harvested by the SSH Open Marketplace, using manually curated files on GitHub as a source. This has been proven challenging for the continuous ingest pipeline. And that is why a more structured workflow is now under investigation, opening up opportunities to also join the (manual) curation efforts of both teams leading to better metadata description of the CRF records.

Second, we will look into improving the connection of the Marketplace with the VLO, which is a vast discovery portal including almost a million metadata records harvested from 47 CLARIN Centres and various non-CLARIN sources like Europeana or ELRA. Indeed, the SSH Open Marketplace by its nature extends beyond the CLARIN world and it could be interesting to investigate in what way the SSH Open Marketplace could complement the VLO, the CLARIN Resource Families and the tools in the Language Resource Switchboard both from a technical point of view (i.e. mutual harvesting) as well as from the points of view of increasing the findability and accessibility of language data.

Third, the SSH Open Marketplace extends CLARIN's quite complex infrastructure of discovery portals, which already includes the VLO, the Language Resource Switchboard and the CLARIN Resource Families. This multitude of discovery portals can be confusing for researchers or developers that would like to include information about their resource into the CLARIN infrastructure and want to ensure the maximum visibility for the community. The same could be the case for those researchers who aim to discover resources. These users currently have at their disposal more and more discovery platforms useful

[30] see https://www.deepl.com/translator

[31] see https://translate.google.com

to access resources as data, tools and services. To diminish the risks of confusion, it is therefore planned to create a guide that clearly outlines the various options of discovery portals, with their similarities and complementarities, to better inform and guide users who want to share their resources, as well as users who search resources. For instance, while the SSH Open Marketplace allows to discover the same language data and tools as the CLARIN discovery portals, it presents a number of differentiation points, as follows:

- The SSH Open Marketplace showcases all its resources, including CLARIN resources, in a contextualised manner, namely via the relations that can be created at the metadata level with other SSH Open Marketplace resources which can be: ”relates to”, ”is related to”, ”documents”, ”is documented by”, ”mentions”, ”is mentioned by”, ”extends”, ”is extended by”).

- Through contextualization, the SSH Open Marketplace enables putting in relation different types of resources, for instance, services and tools can be put in relation with training materials, datasets, publications and/or workflows. This increases the level of informativeness and usability of resources by allowing users to easily finding and accessing related information about the initial resource of interest.

- The SSH Open Marketplace allows searching, in addition to the classical search by keywords, by type of activity, search that draws on the TaDiRAH, the Taxonomy of Digital Research Activities in the Humanities. This increases the probability of finding CLARIN resources, especially for users who are not acquainted with language data or with the CLARIN VLO, the CLARIN Resource Families and the Language Resource Switchboard.

- Having CLARIN resources in the SSH Open Marketplace broadens the type of potential users that may discover and use them through the fact that scholars from social sciences and humanities disciplines, who do not regularly use language data, use the SSH Open Marketplace to discover new resources relevant for their research.

- In opposition to all other discovery engines, the SSH Open Marketplace proposes a unique type of resource, which becomes more and more important when it comes to reproducible research: workflows. This resource type represents an interesting niche that some CLARIN users have already adopted (as explained in the workflows section above) and that will be further promoted among the CLARIN community through a CLARIN Café and hands-on workshops.

Finally, new EOSC-related projects in which CLARIN is involved, such as Open Science Plan-Track-Assess Pathways (OSTrails)[32], will also investigate if and how the SSH Open Marketplace could play a role, alongside other research resource catalogues, in a wider (SSH) knowledge graph. These developments need to be closely monitored, keeping in mind that the curation work currently happening in the SSH Open Marketplace and the records enrichment should continue to serve the research community, no matter which form the resource catalogues will take in the future.

4 Conclusion

As one of the founders of the SSH Open Marketplace, CLARIN ERIC supports the SSH Open Marketplace and its development for the benefits of the CLARIN community and to increase the findability and re-usability of CLARIN resources beyond its target user community. Such a joint enterprise proposing shared services between ERICs has the strong potential to better serve the SSH communities. This is true especially in the EOSC context, where the Cluster approach seems to be an appropriate level to join forces for various scientific domains. Having shared and flagship services, such as the SSH Open Marketplace, for the Cluster demonstrates collaboration capacities between ERICs, while ensuring long-term sustainability, for the benefits of the research community. The process of integrating CLARIN resources into the SSH Open Marketplace exemplifies some of the overall challenges discovery portals such as the

[32] see https://cordis.europa.eu/project/id/101130187

marketplace face. These challenges are generally expressed in the difficult balance between an overly universal and overly specific approach to the data structure. As every new item potentially increases the curation work and requires new, subject-specific knowledge on the side of the curation team, it is important that both the technical and curatorial side are involved in future development of the marketplace. The SSH Open Marketplace can benefit from the CLARIN community expertise in curation work to guarantee metadata quality, especially in aggregation context, based on the knowledge of the VLO and CRF experiences.

References

Barbot, L., Dolinar, M., Gray, E. J., Grisot, C., Illmayer, K., Kurzmeier, M., & McGillivray, B. (2024). Contextualizing research tools &amp; services through workflows in the SSH open marketplace. *Journal of Open Humanities Data*, *10*(1). https://doi.org/10.5334/johd.192

Barbot, L., Moranville, Y., Fischer, F., Petitfils, C., Ďurčo, M., Illmayer, K., Parkoła, T., Wieder, P., & Karampatakis, S. (2021, February). SSHOC D7.1 System Specification - SSH Open Marketplace. https://doi.org/10.5281/zenodo.4558302

Borek, L., Hastik, C., Khramova, V., Illmayer, K., & Geiger, J. D. (2021). Information Organization and Access in Digital Humanities: TaDiRAH Revised, Formalized and FAIR. Universität Regensburg. https://doi.org/10.5283/EPUB.44951
Other Session 5: Knowledge Representation.

Dombrowski, Q. (2014). What Ever Happened to Project Bamboo? *Literary and Linguistic Computing*, *29*(3), 326–339. https://doi.org/10.1093/llc/fqu026

Dombrowski, Q. (2021). The directory paradox. In *People, practice, power: Digital humanities outside the center* (pp. 83–98). University of Minnesota Press. Retrieved April 28, 2023, from http://www.jstor.org/stable/10.5749/j.ctv2782dmw.9

Ďurčo, M., Barbot, L., Illmayer, K., Karampatakis, S., Fischer, F., Moranville, Y., Ocansey, J. T., Probst, S., Kozak, M., Buddenbohm, S., & Yim, S.-B. (2021, December). 7.2 marketplace – implementation. https://doi.org/10.5281/zenodo.5749465

Fišer, D., Lenardič, J., & Erjavec, T. (2018). CLARIN's Key Resource Families. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*. https://aclanthology.org/L18-1210

Gamba, F., Frontini, F., Broeder, D., & Monachini, M. (2022). Language technologies for the creation of multilingual terminologies. lessons learned from the sshoc project. *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 154–163. https://aclanthology.org/2022.lrec-1.17

Marongiu, P., McGillivray, B., & Khan, A. F. (2024). Multilingual Workflows for Semantic Change Research. *Journal of Open Humanities Data*, *10*, 15. https://doi.org/10.5334/johd.179

Petitfils, C., Dumouchel, S., Larrousse, N., Gray, E. J., Barbot, L., Roi, A., Ďurčo, M., Illmayer, K., Buddenbohm, S., & Parkola, T. (2021, October). D7.5 marketplace - governance. https://doi.org/10.5281/zenodo.5608487

Riondet, C., & Romary, L. (2018). The standardization survival kit: For a wider use of metadata standards within arts and humanities. *Archives et Bibliothèques de Belgique-Archief-en Bibliotheekwezen in België*, *106*, 55–62.

Triple Project Consortium. (n.d.). Triple Vocabulary: An SSH multilingual vocabulary based in LCSH [Institution: National Documentation Centre]. https://doi.org/10.12681/semantics.gr/SSH-LCSH

Zinn, C. (2018). The Language Resource Switchboard. *Computational Linguistics*, *44*, 1–13. https://doi.org/10.1162/coli_a_00329