

Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782

Start Date of Project: 01/01/2019

Duration: 40 months

Deliverable 3.5 Report on citation enabled SSH catalogues and SSH citation exploitation

Dissemination Level	PU
Due Date of Deliverable	31/08/2021 (M32)
Actual Submission Date	16/09/2021
Work Package	WP3 - Lifting Technologies and Services into the SSH Cloud
Task	Task 3.4 Making Data Findable by being Citable
Type	Report
Approval Status	Waiting EC approval (V1.0)
Version	V2.0
Number of Pages	p.1 – p.40

Abstract:

This deliverable is a report on Data Citations in SSH. It pertains to SSHOC Task 3.4 under the responsibility of CNRS. The main goal is to evaluate the compatibility of repositories with the recommendations for citation made during Task 3.4.

The information in this document reflects only the author's views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided "as is" without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.



History

Version	Date	Reason	Revised by
0.0	27/05/2021	First Draft of Table of Contents	Nicolas Larrousse (CNRS), Edward J. Gray (CNRS)
0.1	08/07/2021	Table of Contents Reviewed by Partners	All task participants
0.2	31/07/2021	First Draft	Nicolas Larrousse (CNRS) Edward J. Gray (CNRS)
0.3	15/08/2021	Second Draft	Nicolas Larrousse (CNRS) Edward J. Gray (CNRS)
0.4	27/08/2021	Peer review	Olivier Rouchon (CINES) Christina Bornatici (FORS)
0.5	31/08/2021	Address peer review comments	Nicolas Larrousse (CNRS) Edward J. Gray (CNRS)
0.6	03/09/2021	WP Leader Review	Daan Broeder (CLARIN)
0.7	07/09/2021	Address WP Leader comments	Nicolas Larrousse (CNRS) Edward J. Gray (CNRS)
1.0	15/09/2021	Final version	Nicolas Larrousse (CNRS) Edward J. Gray (CNRS)
2.0	06/10/2021	Updated version	Nicolas Larrousse (CNRS) Edward J. Gray (CNRS)

Author List

Organisation	Name	Contact Information
CNRS	Nicolas Larrousse	Nicolas.Larrousse@huma-num.fr
CNRS	Edward Gray	Edward.Gray@huma-num.fr
CLARIN	Daan Broeder	daan.broeder@di.huc.knaw.nl
CNRS/ISTI	Cesare Concordia	cesare.concordia@isti.cnr.it
DARIAH/UGOE	Jan Brase	brase@sub.uni-goettingen.de
LIBER	Athina Papadopoulou	Athina.Papadopoulou@libereurope.org

Table of Contents

1. Introduction.....	7
1.1 General Context	7
1.2 History of Recommendations Developed in Task 3.4.....	9
1.3 Definitions and selection of criteria for tests based on recommendations.....	11
2. Citation Landscape: Qualitative Analysis	13
2.1 SSHOC Dataverse	13
2.2 VCR (Virtual Collection Registry), a SSHOC Component	15
2.3 CoCoon, a CLARIN Center	16
2.4 LINDAT, a CLARIN Center	17
2.5 ADP-Slovenia, a CESSDA Repository.....	17
2.6 DARIAH-DE Data Federation Architecture, a DARIAH Service	18
2.7 NAKALA, a DARIAH Service	20
2.8 RUN - Repositório Institucional da Universidade Nova de Lisboa, a DARIAH Service	21
2.9 RDS - Repozytorium Danych Społecznych, a DARIAH Service	21
3. Citation Landscape: Quantitative Analysis.....	22
3.1 Methods.....	22
3.1.1 Getting metadata from a Landing Page	23
3.1.2 Getting metadata from DOI Registration Agencies	24
3.2 Analysis.....	26
4. Conclusion - Ways Forward in SSH Data Citations	32
5. References	35
List of Figures.....	37
Appendix: List of repositories examined.....	38

Executive Summary

Citation is a pillar for the construction of knowledge. By creating proper citations in a standardized way researchers can constitute a mesh of linked information for various purposes (from credit to reuse). This becomes increasingly important as the SSHOC Task 3.4 team confronts the realities of Social Sciences and Humanities Research in a digital age, when machine actionability takes on a renewed and vital importance.

After conducting an inventory of data citation practices (SSHOC D3.2 “Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning”) and analysing the citation of data in DH¹ 2019 conference abstracts in order to build specifications for the citation prototype, the team discovered a very diverse landscape of data repositories.

As a result, the team developed recommendations for citation in coordination with SSHOC Work Package 2 (*Communication, Dissemination, and Impact*), validated by external reviewers. These recommendations were used to guide a deeper analysis of citation practices in various SSH repositories and how they correspond to these recommendations in order to have a better idea of the current situation. This analysis was carried out in both a quantitative and qualitative fashion.

For the qualitative part, the main goal was to describe, in detail, a selection of repositories representative of the SSH domain. The choice of repositories was made in collaboration with the SSHOC network in order to have a good representation of the very diverse contexts in SSH. This qualitative analysis focused on how these repositories were constructed to provide data citation services in detail.

For the quantitative part, a list of repositories was already established by SSHOC Task 8.2 “Trust & Quality Assurance” and the team took this opportunity to establish synergies and extract a list of repositories to be checked according to defined criteria regarding data citation. The analysis checked 85 repositories from a list of 125 against a set of 7 criteria. In order to facilitate the work, the team used the citation viewer which is part of the prototype mentioned above. The main result of this quick study is that while there are positive signs, especially with respect to the use of landing pages and Persistent Identifiers (PIDs), there is quite a bit of room for improvement as a lot of repositories do not provide machine actionable information. This makes the prototype the Task 3.4 team is currently developing to create actionable citations all the more useful. It also appears from this work that it will be necessary to manually curate some citations in order to enrich them and make them actionable as the information is not always directly available (e.g., a landing page provides a link to a page which contains metadata expressed in another format).

The result of this study will feed the development of the citation prototype developed in Task 3.4 and also liaise with SSHOC WP7 “Creating the SSH Open Marketplace” to integrate citations in SSH Open

¹ Digital Humanities Conference organized by ADHO (See <https://adho.org/conference>)

Marketplace² with a “Cite As” property in the backend and Cite As box in the frontend interface. Another link exists with the very similar work currently being carried out in CLARIN for the Digital Object Gateway.³

Abbreviations and Acronyms

ADHO	Alliance of Digital Humanities Organizations
ADP	Analyze Data & Promote science (Slovenian Social Science Data Repository)
API	Application Programming Interface
ARK	Archival Resource Key
CESSDA	Consortium of European Social Science Data Archives
CLARIN	Common Language Resources and Technology Infrastructure
COCOON	COLlections de CORpus Oraux Numériques
CODATA	COmmittee on DATA
CTS	CoreTrustSeal
DARIAH	The Digital Research Infrastructure for the Arts and Humanities
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
FAIR	Findable Accessible Interoperable Reusable
FAIRDO	FAIR Digital Object Forum

² The Social Sciences and Humanities Open Marketplace, built as part of the Social Sciences and Humanities Open Cloud project (SSHOC), is a discovery portal which pools and contextualises resources for Social Sciences and Humanities research communities: tools, services, training materials, datasets and workflows. Accessible here: <https://marketplace.sshopencloud.eu/> (under construction).

³ Digital Object Gateway is a CLARIN project under development to get more information from a PID (See <https://github.com/clarin-eric/DOGlib>). One of the goals is to enhance CLARIN Virtual Collection Registry (See <https://collections.clarin.eu>)

FDO	FAIR Digital Object
FREYA Project	A project dedicated to extending the infrastructure for persistent identifiers (PIDs) as a core component of open research
HS or Handle	Handle System, a Persistent Identifier (i.e. handles) System provided by CNRI
IIF	International Image Interoperability Framework
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OLAC	Open Language Archives Community
OPENAIRE	Open Access Infrastructure for Research in Europe
PID	Persistent Identifier
PURL	Persistent Uniform Resource Locator
RDA	Research Data Alliance
RDF	Resource Data Framework
RDS	Repozytorium Danych Społecznych (Social Data Repository)
RUN	Repositório da Universidade Nova de Lisboa (Repository of Universidade Nova de Lisboa)
SSHOC	Social Sciences & Humanities Open Cloud
VCR	Virtual Collection Registry

1. Introduction

1.1 General Context

As the Social Sciences and Humanities continue their evolution in a digital age, data are being recognized as key research outputs and research data repositories are being developed simultaneously by actors in many fields to store these valuable data. One key challenge facing researchers in the Social Sciences and Humanities is the need to cite data in this new context.

Beyond its important role in giving credit to the individuals responsible for creating content, citation is a pillar for the construction of knowledge by iteration which can then constitute a mesh of linked information. While citation is a common practice for publications, research data citation is relatively new in the field of Social Sciences and Humanities (SSH).⁴

Citations have been an essential part of serious scientific research for centuries, as scholars seek to prove their claims with evidence and to give credit to the work of fellow scholars. This primordial role of citations has taken on a new, complementary *raison d'être*, namely providing access to the data that one uses. Whereas before a researcher could simply cite the paper or archival document from which they drew their evidence, today's practices increasingly demand that researchers provide access to and allow for the reuse of data. Indeed, this is one of the primary goals of the FAIR initiative, which seeks to render data *Findable, Accessible, Interoperable, and Reusable*.

The main goal of Task 3.4 of the SSHOC Project is to foster the use of data citations by all actors involved in SSH research. One possible way to achieve that goal is to facilitate the use of citations for research projects by providing tools to create standardized citations and make them machine actionable in order to facilitate their dissemination by using automated and standardized protocols. This will give greater visibility to the research data used in Social Science and Humanities following FAIR data principles.

The first action the task undertook was to make an inventory of citation practices. This inventory showed a very diverse landscape within practices in SSH.⁵ In brief, while many initiatives have been proposed, no real standardization for machine actionable citations exists in SSH: there are different communities of practice within different scholarly disciplines. It was clear that it was necessary to adapt the initial description of the task to fulfil this need for standardization. Standardization is the cornerstone of

⁴ Nicolas Larrousse, Daan Broeder, Jan Brase, Cesare Concordia, & Vasso Kalaitzi. (2019). SSHOC D3.2 Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning (Version v1.0). Zenodo. <https://doi.org/10.5281/zenodo.4436736>

⁵ *Ibid.*

automatic processing and dissemination of information which makes machine actionability of a given citation possible. Standardization can be seen at several levels:

- Using a common protocol to exchange information among machines
- The content should be structured in a consistent and comprehensive manner: in this regard the possibility of dealing with more than one format (e.g. From DCTerms to Twitter) can be considered but it should be self-explanatory
- The content itself should use well known “vocabularies” such as GeoNames, OrcID mainly to enrich and liaise information automatically

In order to achieve these objectives, a new work plan was defined. The first step was to create a set of recommendations, based on Force11 principles for data citation, adapted to the specificities of SSH communities. In parallel, the team described the concept of what the team called “FAIR SSH citations” for citations that are FAIR data objects themselves, like FAIR Digital Object⁶ and created a prototype to implement that concept.

Here are some steps to build a FAIR SSH citation with the prototype:

- Take an existing citation (e.g., string) or other types of information (e.g. PID a minima, author etc.)
- Process the string to put it in a standard structure
- Provide access to the Digital Object to which the citation refers.
- Aggregate other information from different sources for instance based on the PID
- Add semantic annotations, whether manually or automatically
- Create a citation viewer
- Provide an API to “disseminate” FAIR citation

Several outreach events were born out of this effort and informed not only the recommendations, but also the FAIR SSH Citation prototype. These outreach efforts were necessary to understand what our communities of practice needed, so that the team could attempt a harmonization of data citation practices that will last. During the November 2020 joint event by SSHOC/FREYA/EOSC-HUB, *Realising the European Open Science Cloud*, SSHOC Task 3.4 organized a session on data citation, where different approaches and experiences related to data citation were discussed by speakers from SSHOC and beyond. Following this event, an article was submitted to the 50th Annual LIBER conference on the need for data citation in the current context of data usage, for instance, computational analysis. Additionally, a Bird of a Feather RDA session, concerning the context of citation, was presented during the 17th RDA plenary conference in Edinburgh. This session began discussions that could result in the creation of an RDA Interest Group that would allow for the continuation of the work done in the SSHOC Task 3.4 after the end of the SSHOC project.

⁶ see <https://fairdo.org/> (accessed Sept 2021)

More specifically within SSHOC, a round table of experts was organized in May 2021 to gather input on current data citation practices from different bodies of experts such as RDA, CODATA, OpenAire etc. and adapt them to the specificity of SSH. In June 2021 a workshop “Data citation in practice”, which aimed to present solutions for efficient data citation from different perspectives, was a good opportunity to get feedback from communities as a result of practical manipulations. This work will be continued with a webinar that will be organized in December. This webinar will present the work of this Task to a broader public and give the SSH community a general view of data citation practices and how they should be done.

These different actions, involving external experts from diverse backgrounds, informed our work in the task and helped us refine the specifications of the FAIR SSH citation prototype as well as inform our Data Citation Recommendations. Our goal is not to create an umpteenth set of recommendations that will be ignored, but to create a durable initiative that will lead to better and more comprehensive data citation practices in the SSH communities. After consultation with SSHOC Work Package 7, responsible for the development of the SSH Open Marketplace, it has been agreed that the Marketplace will integrate a “Cite As” feature, alongside metadata embedded in the source code of the web page, which will continue with the best practices for data citation.

This deliverable examines the current state of data citation within repositories, with detailed analysis of a wide variety of research data repositories. This builds off of the previous deliverable, which was about practices of researchers and data creators.⁷ This deliverable begins with an explanation of the “Recommendations for FAIR Data Citation in the Social Sciences and Humanities”⁸ mentioned in the introduction, as the work and analysis carried out are based on these recommendations. They were developed based on consultations with existing recommendations, data citation experts and the various communities of practice. Next, the definition and selection criteria for evaluating current data practices are outlined. Then, using these criteria, this deliverable describes various SSH data repositories and evaluates how their current data practices line up with these recommendations. To do this analysis, the team looked at SSHOC-affiliated repositories, the specific case of CLARIN B and C centres, and other SSHOC affiliated centres. Finally, this deliverable considers some of the ways forward for data citation in the social sciences and humanities.

1.2 History of Recommendations Developed in Task 3.4

As a result of the research for the “Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning” the need to further define recommendations for FAIR Data Citation - especially in the Social Sciences and Humanities - was surfaced. The need for a dynamic

⁷ *Ibid.*

⁸ Nicolas Larrousse, & Edward J. Gray. (2021). Recommendations for FAIR Data Citation in the Social Sciences and Humanities. Zenodo. <https://doi.org/10.5281/zenodo.5361718>

document that could serve as a guide for all important stakeholders dealing with data emerged, especially as it would permit these stakeholders to stay up to date with the developments in data citation within the SSH.

The result was the publication of the “Recommendations for FAIR Data Citation in the Social Sciences and Humanities.”⁹ The set of recommendations was drafted bearing in mind the actors and users of research data within SSH, mentioned at the beginning of the recommendations, and aiming to determine the current challenges they face, their actionable recommendations, and the expected outcomes that they would ideally have. The first draft was validated by reviewers on behalf of several communities that corresponded in the actor and user personas identified. Apart from the initial set of guiding principles proposed in the conclusion of the D3.2 Deliverable, inspiration for drafting the update was also drawn from the workshop organised during the Realising EOSC Conference¹⁰ on “FAIR Data Citation for Social Sciences and Humanities”¹¹. Discussions during the workshop served as the basis to define the update on the recommendations but also validated the need for a dynamic document that could be updated periodically to cater to the needs of the SSH community when it comes to FAIR data citation.

The updated document was drafted by exploring several applicable use cases and by using the FORCE11 principles as the guide to face the societal and technical challenges associated with FAIR data citation in SSH. Then, a set of recommendations was drafted to respond to the challenges identified and to predict their expected outcome. The draft document was then circulated to an expert group of reviewers including Data Stewards, Publishing Managers, Data Service Experts and Researchers to test its relevance and validity. After assessing and incorporating the feedback received from the invited group of experts, the recommendations were finalised, and the updated document was circulated to the participants of the round table of experts on “Data Citation Practices in the Social Sciences and Humanities (SSH)”¹². Three main outputs resulted from that round table: the need for a rewards system for all data creators, the importance of data publication (in data papers, data journals, etc.) and finally the need to draw inspiration from fields that are currently in the forefront of data publication, such as astronomy, to serve as leading examples of data sharing practices.

The vision for the Recommendations for SSH Citation Practices is that they can serve as a dynamic and living document which would be periodically reviewed by experts within the SSH community so that it remains relevant and helpful. The need for guidance and best practices voiced by professionals within

⁹ Nicolas Larrousse, & Edward J. Gray. (2021). Recommendations for FAIR Data Citation in the Social Sciences and Humanities. Zenodo. <https://doi.org/10.5281/zenodo.5361718>

¹⁰ Realising EOSC conference: <https://www.sshopencloud.eu/news/realising-eosc-virtual-conference-difference> (accessed Sept 2021)

¹¹ Fair data citation session: <https://www.eosc-hub.eu/events/realising-european-open-science-cloud/fair-data-citation-ssh> (accessed Sept 2021)

¹² Roundtable of Experts on Data Citation: <https://sshopencloud.eu/news/roundtable-experts-data-citation> (accessed Sept 2021)

the SSH realm points to the necessity of a living document that can serve as a guide for any user and creator of data. Developments in the field of data citation have been coming in quite rapidly and forcefully so far and maintaining the recommendations might be challenging. However, they have the potential to serve as the foundation for any kind of user of data within SSH. It is undeniable that the creation of data and its curation is currently on the rise, so it is important for organisations issuing and supporting the FAIR data citation - in SSH or in any other field - to maintain a dynamic approach and follow the rapid developments. This will ensure their ability to better support the communities they serve.

1.3 Definitions and selection of criteria for tests based on recommendations

This section outlines the definition and selection criteria which are used in the following qualitative and quantitative analyses of data repositories in SSH. The methodology for each analysis appears before each section.

The use of repositories is essential to ensure the implementation of good practices regarding the management of research data. In order to check how different repositories, offer services regarding citations, the team defined a set of criteria based on experience gained during the development of actions in the framework of the task.

This view is based on an extraction of the set of recommendations built in Task 3.4. More precisely, the team took the general principles and examined whether each recommendation was fulfilled for a selection of repositories when it is applicable.

Figure 01: Extraction of a recommendation from "Importance" principles which is applicable to a repository

<u>Societal/Technical Challenge (adapted from FORCE11 principles)</u>	<u>Recommendation</u>
Importance: Current lack of a "data citation culture," "disappearance" of research material by non publication.	Provide a ready to use "cite as" recommendation in one's published work or datasets

Main Criteria from recommendations (checked in the quantitative analysis)

- **PID** from **Unique Identification & Persistence**
 Research data should persist beyond the research project itself: the systematic use of PIDs whatever the technology makes research data easier to find and cite.

- **Landing page** from **Access**
Provide the most comprehensive metadata possible in a landing page, such as extra documentation to foster their possible reuse.
- **Structured metadata Importance & Credit and Attribution**
Metadata should be machine readable and in order to be processed, they should be structured in a common format such as DCterms, Schema.org etc.
- **Cite as** from **Evidence, Specificity & Verifiability**
A ready to use citation raises awareness of the contributor's work for stakeholders, such as funding agencies, publishers, universities and research institutes. It also ensures that all people involved in the creation of research data receive proper credit.
- **Versioning** from **Specificity and Verifiability**
The main goal is to enhance trust in the process of creation of the data set
- **Standardized vocabularies** from **Interoperability and Flexibility**
Structured metadata are not sufficient, the content should also be standardized in order to be processed: for instance, use of ORCID for persons, Creative Commons for licences etc.
- **Links to publications** from **Importance**
It is an important way to give more visibility to a data set and also favour the reproducibility of research. This is also the foundation for a graph which links data, publications, authors, institutions etc.

During the evaluation, it was decided that PID, Landing Page, Structured Metadata and Cite As were more important and consequential for a robust and well-structured citation environment, as these are the basic components upon which machine actionable data citation is founded. Standardized vocabularies, versioning, and links to publications, while important, were not quite as fundamental in our analysis.

Other Criteria from recommendations (not checked in the quantitative analysis)

- **Granularity** from **Specificity and Verifiability**
It is an important specificity of SSH data in general. There is a need to be able to cite a whole dataset, a record or even a part of record (e.g. for audio recordings)
- **Citation counter** from **Credit and Attribution**
This is a complement to the link that must be established between data and publications (but not only) that was mentioned above
- **Metadata harvested by an aggregator** from **Importance, Credit and Attribution, Persistence**
It is not always easy to verify this criterion but it is also a very important step to give broad visibility to both dataset and stakeholders

These criteria were generally more difficult to test and due to the low amount of uptake observed in the sample, they would not have greatly impacted the analysis or distorted the results.

2. Citation Landscape: Qualitative Analysis

This section considers the current state of citation creation and use especially as provided by repositories and infrastructure components currently used in SSH. Some of these components are currently being created or worked on in the SSHOC project such as the SSHOC Dataverse and Virtual Collection Registry. Others have a longer history, and are independently operated or associated with a research infrastructure such as CLARIN.¹³ For instance during the SSHOC workshop “Citation in practice”, solutions were presented for efficient data citation from different perspectives from the CLARIN world (VCR Virtual Collection Registry, LINDAT and Cocoon).

The selection of these repositories was conducted through the SSHOC network for instance repositories that asked for support in the process of Core Trust Seal certification as provided by SSHOC Task 8.2 “Trust & Quality Assurance.” The first two entries, the SSHOC Dataverse and the Virtual Collection Registry (VCR), are being developed internally as part of the SSHOC project. The other repositories were selected from SSHOC network of partners and are intended to represent a variety of disciplines and the various ERICs involved in SSHOC such as CESSDA, CLARIN, and DARIAH. These choices may seem arbitrary, but they come from those known to the SSHOC project and aim at giving a synthetic vision of current practices regarding the citation of data in SSH.

These deeper, qualitative analyses are coupled with a quantitative overview in the next section, for which the team used the inventory of SSH repositories created in SSHOC task 8.2, who compiled this for an investigation of potential CoreTrustSeal (CTS) certification subjects. But mainly the team leveraged the expertise of D3.5 contributors with regards to their respective infrastructure repositories and components.

2.1 SSHOC Dataverse

Dataverse is an open source software platform for storing, sharing, citing, and preserving research data, originally designed and developed by the Data Science and Products team¹⁴ at the Institute for Quantitative Social Science (IQSS) of Harvard University, and is maintained and improved by a large community of contributors forming the Dataverse community¹⁵. The Dataverse platform provides several functionalities to enable users to share and publish their data, in particular, considering the mandatory criteria defined in section 1.3:

¹³ CLARIN is based on a distributed network of technical centers. The main backbone is made of centers that provide services (B-Centers) to the CLARIN community. There are currently 20 B-centers. Other centers provide metadata (C-Centers) that are integrated with CLARIN but they need not offer any further services. CLARIN provides general recommendations adapted to specific needs of language resources.

¹⁴ <http://www.iq.harvard.edu/people/people/data-science-products> (accessed Sept 2021)

¹⁵ <https://dataverse.org/developers> (accessed Sept 2021)

- it provides a function that automatically generates the citation string when a dataset is created by a user. The citation string¹⁶ is built according to recommendations defined in the Joint Declaration of Data Citation Principles (2014)¹⁷,
- it can be configured to automatically generate a unique global identifier, that can be a Handle or a DOI, for the dataset. The identifier gives access to a landing page and furthermore to the digital object
- it enables users to access and export the metadata of the dataset using the 'Metadata' tab present in the landing page. The metadata can be exported in several formats: Dublin Core, DDI, Data Cite 4, Json etc

Additionally, the Dataverse platform creates a new version of the dataset each time an update is made to one of the digital objects forming it; all versions of a dataset can be accessed by clicking on the 'Version' tab of the landing page, and for every version the list of changes is reported. To make sure that a citation string refers to a specific version of the dataset, a Universal Numerical Fingerprint (UNF) is generated for every version and added to the citation string itself. The UNF is a string that unequivocally identifies a version of a dataset, ensuring researchers that they are referencing a particular version of the dataset.

Figure 02: Example of a citation string generated by a Dataverse platform

Hanmer, Michael J.; Banks, Antoine J., White, Ismail K., 2013, "Replication data for: Experiments to Reduce the Over-reporting of Voting: A Pipeline to the Truth", Harvard Dataverse, V1, <http://dx.doi.org/10.7910/DVN/22893> UNF:5:eJOVAjDU0E0jzSQ2bRCg9g==

There are currently many organizations in several countries officially using the Dataverse platform to publish their research data.¹⁸

Technically, the Dataverse platform is based on a modular design principle and provides APIs. This enables developers to easily integrate the platform in other systems and to extend its functionalities by building microservices on top. Using these features, specifications for a Dataverse platform have been produced in the SSHOC project (WP 5.2) with specific functionalities targeted to SSH organizations. The main features of SSHOC Dataverse are: multilingual web interface, localized metadata fields, support for data standardization techniques based on APIs for CESSDA CVs, Topic Classification and CESSDA CV Manager services, integration with the Edugain/SURFconext federated login infrastructure etc. The list of

¹⁶ <https://dataverse.org/best-practices/data-citation> (accessed Sept 2021)

¹⁷ <https://www.force11.org/datacitation> (accessed Sept 2021)

¹⁸ The complete list can be found at <https://dataverse.org/installations> (accessed Sept 2021)

features implemented in SSHOC Dataverse, alongside with the source code, can be found on the SSHOC GitHub repository.¹⁹

The SSHOC Dataverse platform will be provided to those SSH organizations that need a repository system to store and manage research data. SSH organizations with limited technical resources could use it as a cloud service to build and manage an online repository, while organizations already providing archival solutions can use it to set up a sharing and self-depositing environment for researchers in a user-centric manner. The SSHOC Dataverse repository is not yet in use, but it will offer functionalities that reflect the latest insights (e.g. related to FAIR) developed in the context of SSHOC project.²⁰

2.2 VCR (Virtual Collection Registry), a SSHOC Component

The CLARIN Virtual Collection Registry²¹ is a service that enables users to create and register virtual collections, which are sets of references to resources from different repositories, together with some descriptive metadata describing the collection. The CLARIN VCR is already a production quality service that is registered in the EOSC service catalogue.²² SSHOC task 3.6 “Making data Reusable and actionable” discussed with interested SSHOC partners representing the broad Social Sciences and Humanities community how to make the VCR more suitable and appealing for users of the partner infrastructures.

The main purpose of Virtual Collections (VC) is to allow users to bookmark and manage sets of distributed and heterogeneous data as a single unit. This also includes the use of citing VCs in papers or other research products.

The VCR currently implements the following citation functionality:

- generation of a PID both a DOI and EPIC Handle for identifying the VC, although only the DOI is used in the citation text generated
- generation of BibTeX citation text
- citation text content: Title, Authors, Year, Link (actionable PID)

This is in line with the CLARIN citation policy where CLARIN supports²³ Force11 Data Citation Principles. BibTeX is a popular format that is, for example, also provided by LINDAT and ARCHE CLARIN centres.

With respect to VCR citation idiosyncrasies there should be mention of the particular version policy the VCR implements. Usually VC cited datasets are static, however the VCR makes it possible to adapt a VC

¹⁹ <https://github.com/SSHOC> (accessed Sept 2021)

²⁰ the main SSHOC partner (DANS) responsible for SSHOC Dataverse development is also deeply involved in the FAIR initiative and projects

²¹ <https://collections.clarin.eu/> (accessed Sept 2021)

²² <https://marketplace.eosc-portal.eu/services/virtual-collection-registry/details> (accessed Sept 2021)

²³ <https://www.clarin.eu/content/persistent-identifiers> (accessed Sept 2021)

for instance by adding more components. This is not indicated in the citation text, while (informal) CLARIN PID usage recommendations recommend making a particular versioning policy explicit, which is also a CLARIN-B centres requirement. It is expected that this will be resolved in a future release.

2.3 CoCoon²⁴, a CLARIN Center

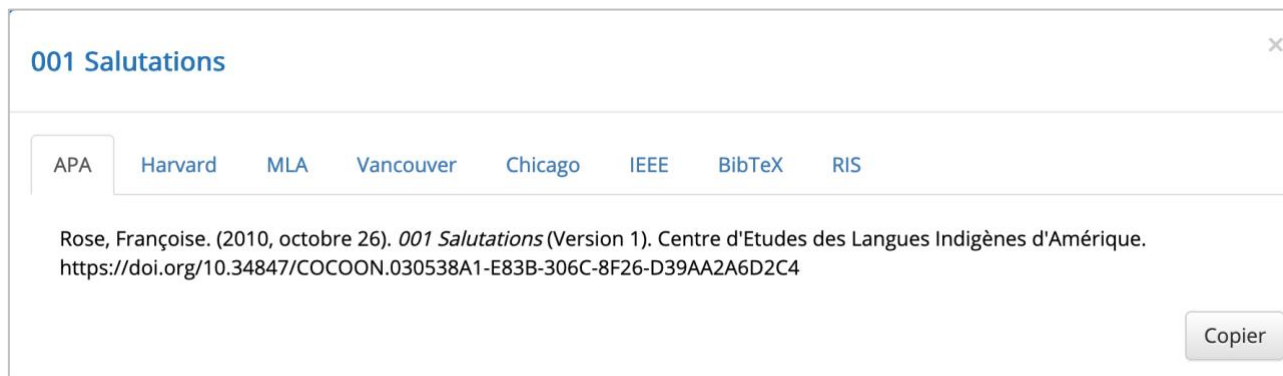
Cocoon is a French repository dedicated to oral data. CoCoon hosts recordings (audio & video) and associated annotations (Transcriptions, Translations, Measurements etc.). Descriptive metadata are expressed in OLAC²⁵ format. An extended curation of resources is performed on resources before publication (e.g., file formats). CoCoon contains around 14000 recordings expressed in 250 languages, totalling approximately 5000 hours of recordings. All records are deposited on the CINES²⁶ platform for long term preservation. The content is harvested by OpenAire, ISIDORE and the specialized OLAC Portal.

CoCoon uses three types of PIDs for different purposes (e.g. DOI to access the landing page of the data, PURL as OAI²⁷ identifier, ARK for preserved items).

Metadata are made accessible through different technologies. Actionable metadata, expressed in different formats are embedded in the landing page of a resource.

CoCoon also provides different styles of citation for a resource.

Figure 03: Metadata from CoCoon Landing Page expressed in various citation styles



001 Salutations ×

APA
 Harvard
 MLA
 Vancouver
 Chicago
 IEEE
 BibTeX
 RIS

Rose, Françoise. (2010, octobre 26). *001 Salutations* (Version 1). Centre d'Etudes des Langues Indigènes d'Amérique. <https://doi.org/10.34847/COCOON.030538A1-E83B-306C-8F26-D39AA2A6D2C4>

As CoCoon deals with audio recordings, it provides a possibility to cite part of a record based on the W3C standard, "Media Fragments URI."²⁸

²⁴ <https://cocoon.huma-num.fr> (accessed Sept 2021)

²⁵ <http://www.language-archives.org/> (accessed Sept 2021)

²⁶ <https://www.cines.fr/en/long-term-preservation/> (accessed Sept 2021)

²⁷ https://en.wikipedia.org/wiki/Open_Archives_Initiative_Protocol_for_Metadata_Harvesting (accessed Sept 2021)

²⁸ <https://www.w3.org/TR/media-frag/> (accessed Sept 2021)

Metadata from CoCoon are also disseminated through OAI-PMH protocol (e.g. with specific OLAC format) and are also available in a RDF TripleStore with a SPARQL EndPoint.

2.4 LINDAT²⁹, a CLARIN Center

The LINDAT CLARIN Centre for Language Research Infrastructure provides technical background and assistance to institutions or researchers who want to share, create and modernise their tools and data used for research in linguistics or related research fields. The project also provides an open digital repository and archive open to all academics who want their work to be preserved, promoted and made widely available. LINDAT is based on D-Space technology which is very popular and used by around 10 CLARIN centres. Regarding Data Citation, LINDAT provides metadata to various metadata service providers in specific formats (CLARIN, Google data search, OpenAire, EUDAT, Data Citation Index etc.). In order to do so, metadata are embedded in the landing page (e.g. DCTerms, schema.org etc.). LINDAT provides ready to use “Cite As” in various formats (simple strings, BibTeX, CMDI etc.). There are plans to add support for the Citeproc³⁰ JSON format which is used by Zotero³¹, Mendeley³² and others.

2.5 ADP-Slovenia³³, a CESSDA Repository

ADP is a repository dedicated to Social Science Data and a member of the CESSDA community. Founded in 1997 at the Social Sciences Research Institute at the Faculty of Social Sciences, University of Ljubljana, it is the Slovenian national infrastructure for social sciences. Its mission is thus “to ensure and promote sustainable services of ingest, storage and access to quality research data from the field of Slovenian social sciences and broader.”³⁴ ADP provides a series of pedagogical materials, such as the importance of Open Data, the research data lifecycle. It proposes an analysis of a dataset based on Nesstar software.

ADP provides a productive citation environment. It provides a dialogue box indicating a citation recommendation, clear Terms of Use, a landing page with metadata using schema.org. It provides for a DOI to be linked, but not every entry has a DOI, due to the source from which it was ingested lacking this persistent identifier. It can account for which version of the dataset is linked. It has an entire tab related to “Accompanying materials” such as the materials used by the study, those linked to results, and related publications. ADP is listed in re3data.org and is CTS certified.

²⁹ <https://lindat.mff.cuni.cz/en> (accessed Sept 2021)

³⁰ <https://en.wikipedia.org/wiki/CiteProc> (accessed Sept 2021)

³¹ <https://www.zotero.org/> (accessed Sept 2021)

³² <https://www.mendeley.com> (accessed Sept 2021)

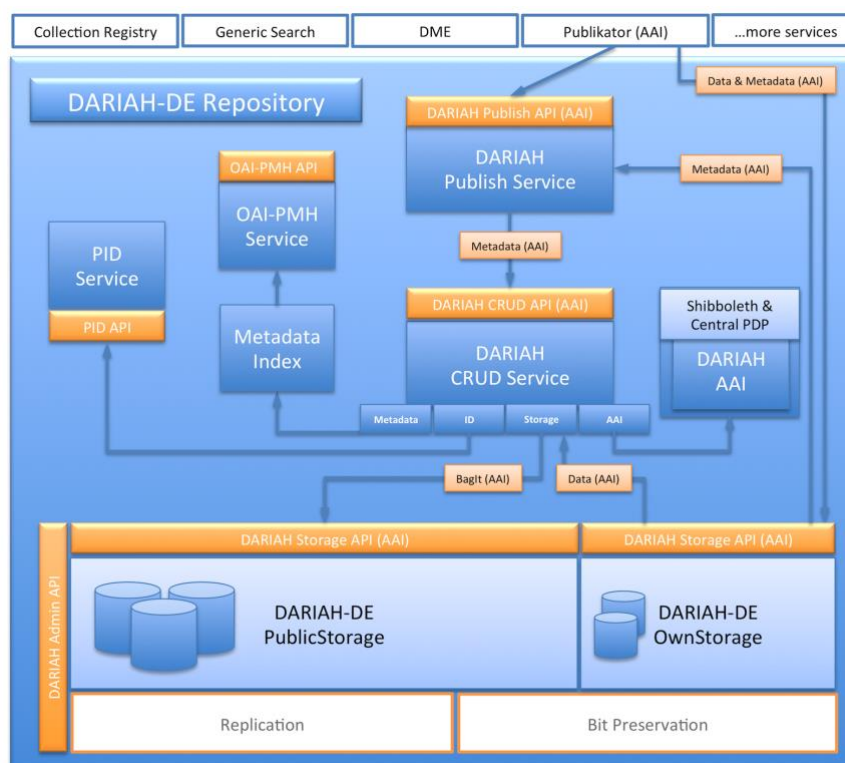
³³ <https://www.adp.fdv.uni-lj.si/> (accessed Sept 2021)

³⁴ <https://www.adp.fdv.uni-lj.si/eng/spoznaj/adp/poslanstvo/> (accessed Sept 2021)

2.6 DARIAH-DE Data Federation Architecture, a DARIAH Service

DARIAH stands for “Digital Research Infrastructure for the Arts and Humanities” and supports research in the humanities and cultural sciences with digital methods and procedures. DARIAH-EU is a European Research *Infrastructure* Consortium (*ERIC*), with DARIAH-DE as the German members, operated by the University of Göttingen. DARIAH-DE operates a repository as a digital long-term archive for human and cultural-scientific research data. The DARIAH-DE Repository is a central component of the DARIAH-DE Research Data Federation Architecture, which aggregates various services and applications for easy use by DARIAH users. The DARIAH-DE Repository allows researchers to save their research data in a sustainable and secure way, to provide it with metadata and to publish it. Any uploaded collection as well as each individual file is available in the DARIAH-DE Repository in the long term and is assigned two unique PIDs: a DataCite DOI³⁵ for citation, and an EPIC Handle PID³⁶ for administrative use. In addition, researchers can register their collections within the Collection Registry, which are then also found in the Generic Search.

Figure 04: The DARIAH-DE Repository Architecture



³⁵ <https://datacite.org/> (accessed Sept 2021)

³⁶ <https://de.dariah.eu/pid-service> (accessed Sept 2021)

DataCite DOI and EPIC Handle prefixes are institution specific, and the suffixes for DARIAH-DE Repository DOIs and Handles are just identical, such as:

- 10.20375/0000-000B-C8EF-7 (DataCite DOI)
- 21.11113/0000-000B-C8EF-7 (EPIC Handle)

Each object is stored as a Bagit³⁷ bag in the DARIAH-DE PublicStorage, where it can be accessed publicly. A Bagit profile for the DARIAH-DE Repository Bags is provided in Version 0.1.³⁸

Access to the repository's objects is provided using HTTP content negotiation with the basic DOI or Handle. You can get:

1. The complete bag (as ZIP) setting HTTP's Accept-Header to application/zip.
2. The HTML landing page if requesting text/html.
3. The data object itself otherwise.

Researchers can access the repository's content using the DOI, the Handle, and directly via the DH-crud URL. All objects in the DARIAH repository can be cited by using the DOI and additional metadata required by the citation style used.³⁹ For an automated citation process, the metadata can be retrieved automatically. Researchers can furthermore retrieve the descriptive RDF metadata of each object directly in various formats, by adding @metadata/[format] to the handle followed by the preferred format:

- <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@metadata> (same as <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@metadata/ttl>)
- <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@metadata/xml>
- <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@metadata/json>
- <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@metadata/jsonld>
- <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@metadata/ntriples>

It is also possible to retrieve administrative RDF metadata in various formats by adding @adm/[format], following the same schema.

Technical metadata is provided in FITS XML ([File Information Toolset](#)) only, see [FITS XML schema](#):

- <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@tech>

The landing page provides basic information about the object, and leads to easy data download and basic metadata.

³⁷ Bagit: <https://tools.ietf.org/html/draft-kunze-bagit> (accessed Sept 2021)

³⁸ Available here: https://repository.de.dariah.eu/schemas/bagit/profiles/dhrep_0.1.json (accessed Sept 2021)

³⁹ For instance, by using a Citation Formatter <https://citation.crosscite.org/> (accessed Sept 2021)

- <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@landing>

The index page provides a more technical view of the object, and includes links to all metadata, formats, and to Handle metadata.

- <https://hdl.handle.net/21.11113/0000-000B-C8EF-7@index>

Additionally, all DOI and Handle URLs are directly resolved to the DARIAH-DE CRUD service, which unpacks the BagIt ZIPs from PublicStorage and delivers all the information. An example for all metadata information can be found at:

- <https://repository.de.dariah.eu/1.0/dhcrud/21.11113/0000-000B-C8EF-7/metadata>.

2.7 NAKALA⁴⁰, a DARIAH Service

NAKALA is a French interoperable and secure service for depositing all types of data (e.g. text files, audio, video, images or other types) in order to share them. This repository mainly provides the following services:

- assignation of a PID (Persistent IDentifier) making data and metadata citable;
- permanent data access;
- dissemination of metadata through a Triple Store and OAI-PMH;
- dedicated search engine;
- customized presentation with NAKALA Press.

This allows the separation of data management from data presentation. NAKALA now uses DOIs as Persistent IDentifiers and previously used Handles. Actionable metadata are embedded into the landing page in various formats, such as Dublin Core, OpenGraph, Twitter, schema.org. It provides a citation string on the landing page. Metadata are also disseminated through OAI-PMH protocol and are also available in a RDF TripleStore with a SPARQL EndPoint.

Figure 05: Citation extracted from the landing page of DOI 10.34847/nkl.7847b549

Citer

Jarry, M. (2021) «Site_0145_Fig.92_3. Céramique de l'Âge du Bronze» [Image] NAKALA. <https://doi.org/10.34847/nkl.7847b549>

⁴⁰ <http://nakala.fr> (accessed Sept 2021)

NAKALA is harvested by various aggregators, ISIDORE from Huma-Num, Gallica from the French National Library and OpenAire. NAKALA proposes a REST API to classically manage data and query the repository and provide an access to vocabularies used internally⁴¹. NAKALA is developed and maintained by Huma-Num, the French infrastructure dedicated to providing services to SSH communities.

2.8 RUN - Repositório Institucional da Universidade Nova de Lisboa⁴², a DARIAH Service

RUN is the data repository for the Universidade Nova de Lisboa. It aims to collect, store, manage, preserve, and give access to the university's intellectual production, serving professors, researchers, students, alumni, and collaborators in NOVA projects.⁴³ All documents available on the RUN repository are available via CC-BY-NC. RUN references over 33,329 publications and had over 2,129,937 downloads in 2020 alongside 723,993 queries.⁴⁴ Authors upload their work, and the metadata are verified by the documentalists and librarians in charge of the repository. The metadata of RUN is included and searchable in the RCAAP⁴⁵ (Open Access Scientific Repository of Portugal) and in the ROSSIO Infrastructure. Additionally, RUN adheres to the guidelines of the European DRIVER project. It is featured in OpenAire Explore.

Items in the RUN Repository are given a handle, and users are instructed to use this PID to cite or link to a given item. The metadata follows the Dublin Core Metadata Terms and is encoded in their website in such a way as to be machine actionable, using the OAI-PMH protocol, as shown with Zotero or the FAIR SSH Data Citation Prototype.

2.9 RDS - Repozytorium Danych Społecznych⁴⁶, a DARIAH Service

The RDS (Repozytorium Danych Społecznych – Social Data Repository) is a repository archiving social data. It includes data from PADS (Polskie Archiwum Danych Społecznych – Polish Social Data Archive, a joint enterprise of the Robert Zajonc Institute for Social Studies, University of Warsaw and the Institute of Philosophy and Sociology of the Polish Academy of Sciences) and ADJ (Archiwum Danych Jakościowych

⁴¹ See <https://api.nakala.fr/doc> (accessed Sept 2021)

⁴² <https://run.unl.pt> (accessed Sept 2021)

⁴³ Policy document of the RUN:

<https://www.biblioteca.fct.unl.pt/sites/www.biblioteca.fct.unl.pt/files/documents/pdf/run.policy.pt.pdf> (accessed Sept 2021)

⁴⁴ Available from RUN stats: <https://run.unl.pt/stats?level=general&type=access&page=downviews-series> (accessed Sept 2021)

⁴⁵ See <https://www.rcaap.pt/> (accessed Sept 2021)

⁴⁶ <https://rds.icm.edu.pl/> (accessed Sept 2021)

– Qualitative Data Archive, run by the Institute of Philosophy and Sociology of the Polish Academy of Sciences). Launched in 2020, in August 2021 the repository included more than 250 datasets and over 850 files, with almost two thousand downloads. The PADS published a lengthy “Social Data Preparation Manual” to help scholars ensure that their data are correctly formatted and ready for archiving, though this is available only in Polish.⁴⁷ The repository is featured in OpenAire Explore.

RDS provides a robust data citation environment. The repository applies the FAIR principles, with each resource described with standardized metadata and a DOI. RDS provides a suggested citation component, which allows users to easily cite the dataset and download citation data in EndNote XML, RIS, and BibTeX formats.

RDS provides for versioning and metadata exportation in a variety of formats (Schema.org JSON-LD, OAI_ORE, JSON, OpenAIRE, DDI, DataCite, Dublin Core). Users are able to include information about related publications, as well as granularly assign licenses to individual files. It is also possible to approach a dataset contact person via the contact dialog box.

The operator of RDS is the Interdisciplinary Centre for Mathematical and Computational Modelling, University of Warsaw. The repository operates via a modified version of the Dataverse software.

3. Citation Landscape: Quantitative Analysis

3.1 Methods

In addition to the qualitative review of repositories in the previous section, a quantitative analysis was also conducted in order to get a broader and more profound understanding of the data citation landscape of the SSH. A list of repositories was provided from Task 8.2, which is also examining the status and functionality of SSH data repositories.⁴⁸ Synergizing with their work, the team evaluated how well these repositories followed the most important elements of the data citation recommendations.

The original list contained 125 repositories, of which 85 were able to be checked. The rest were not included for various reasons:

- the site did not really contain research data (e.g. description of a project or search engine into ancient texts or simply an encyclopedia)
- the data were outside the scope of SSH even if the overall project was related (e.g. database of biological life science records)

⁴⁷ <https://pads.org.pl/wp-content/uploads/2020/05/podrecznikADS.pdf> (accessed Sept 2021)

⁴⁸ See list in Appendix.

- the site was not functional
- there were some redundancies (the same repository was listed multiple times).

In the process of examining these 85 repositories, the team was unable to do an in-depth, exhaustive examination, as was performed in the previous section. Nevertheless, the team considers that this broad ranging, quantitative approach bears significant fruit for understanding the current data citation landscape in SSH. The team was also limited by certain access problems as well as language barriers to what should be considered to still be informative data, which demonstrates the importance of writing one's repository to be accessible to all and maintaining it.

In our evaluation, the team examined:

- Does the repository provide a persistent identifier (PID)? And if so, what is the typology of this PID (e.g. DOI, handle, URI, URL etc.)?
- Does the PID give access to a landing page and furthermore to the digital object?
- Are structured metadata available through the landing page or via another standardized protocol (API, OAI-PMH, TripleStore etc.)?
- Does the repository itself provide a ready to use "cite as" citation formatter? Or alternatively, does the technology used for the PID have a service to do it (e.g. DOI citation formatter <https://citation.crosscite.org/>)?
- Does the repository account for versioning?
- Does the repository use "standardized" vocabularies (e.g. ORCID for attribution)?
- Does the repository provide links to publications (e.g. Google Dataset Search)?

The analysis was conducted through both a manual analysis of the landing pages, as well as an automatic analysis of DOI Registration Agencies with the Data Citation Prototype. These two ways of gathering metadata are described below. The citation viewer from the FAIR citation prototype, described in the introduction, implements these different strategies to retrieve metadata.

3.1.1 Getting metadata from a Landing Page

From this survey, it's clear that nearly all repositories provide landing pages even if these landing pages have different levels of sophistication.

This process is essentially manual: the idea is to see if it is possible to retrieve information from the HTML source code of the landing page. There are several ways to embed metadata in a landing page. It's possible to use the "meta" tag (e.g. `<meta name="dc:identifier" content=" ...">`) for DublinCore, OpenGraph and other formats.

Alternatively formats like "schema.org" use a script to embed metadata in Json (e.g. `"@context": "http://schema.org", "@id": "MyDOI", "@type": "Dataset" etc.`)

Process of verification from a landing page

For all the repositories checked for this deliverable (See repository list in the Appendix), the analysis checked all mandatory criteria. Even if some criteria can be verified automatically, it is necessary to perform most of the checks manually.

Anyway, in order to get a quick overview of what can be grabbed from a specific resource hosted by a repository, the team used the citation viewer⁴⁹, developed in the task for the FAIR SSH citation prototype, to retrieve metadata.

Figure 06: Sample of table used to check repositories

ID from task 8.2	Repository Name	URL	PID	Landing Page	Structured metadata	Cite As	Versioning	Standardized Vocabularies	Link to Publication
999	MyRepo	https://myrepo.org	DOI	Yes	schema.org	Yes	No	No	Yes

3.1.2 Getting metadata from DOI Registration Agencies

The DOI Registration Agencies (RA) store metadata about items identified by DOIs they resolve and publish and distribute metadata in two ways: by using HTTP Content Negotiation⁵⁰ and by providing API entries to query the metadata repository.

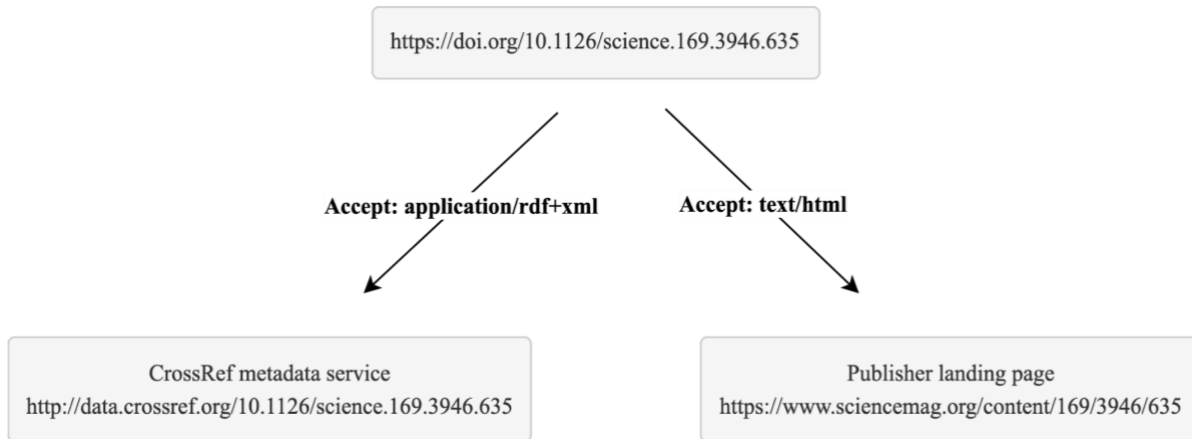
The approach adopted for HTTP Content Negotiation is the ‘server-driven’ one.⁵¹ In this approach HTTP clients and servers negotiate a possible answer to a specific request using HTTP headers. When a client requests a resource using a DOI, the RA server checks the expected media type; if the client expects an RDF media type the request URI is redirected to the RA metadata server, which provides the metadata record related to the DOI, formatted accordingly.

⁴⁹ Available here: <http://v4e-lab.isti.cnr.it/citview/demo/> (accessed Sep 2021)

⁵⁰ RFC 2616 – Hypertext Transfer Protocol – HTTP/1.1 – (Section 12: Content Negotiation)

⁵¹ <https://citation.crosscite.org/docs.html> (accessed Sep 2021)

Figure 07: Server-driven HTTP Content Negotiation



The content negotiation then enables users to retrieve all metadata related to a digital object, starting from the DOI of the object.

The API published by RAs enables users to execute more complex interactions, compared to content negotiation, for instance the API enables users: to search for metadata using bibliographical references other than DOIs, to retrieve in one request several metadata records, or to request only specific metadata values etc. For instance, with Crossref API⁵² the metadata repository can be queried using strings that contain bibliography references, the string does not have to be a well-written citation string. The string is parsed using machine learning techniques and the Crossref system tries to match the reference string with the metadata that are stored in the database. It will return a metadata record and a score indicating the probability that the returned record is the correct one.

Technically, RAs API are published as RESTful APIs developed according to the JSONAPI specifications. API documentations of the main DOI RAs are available:

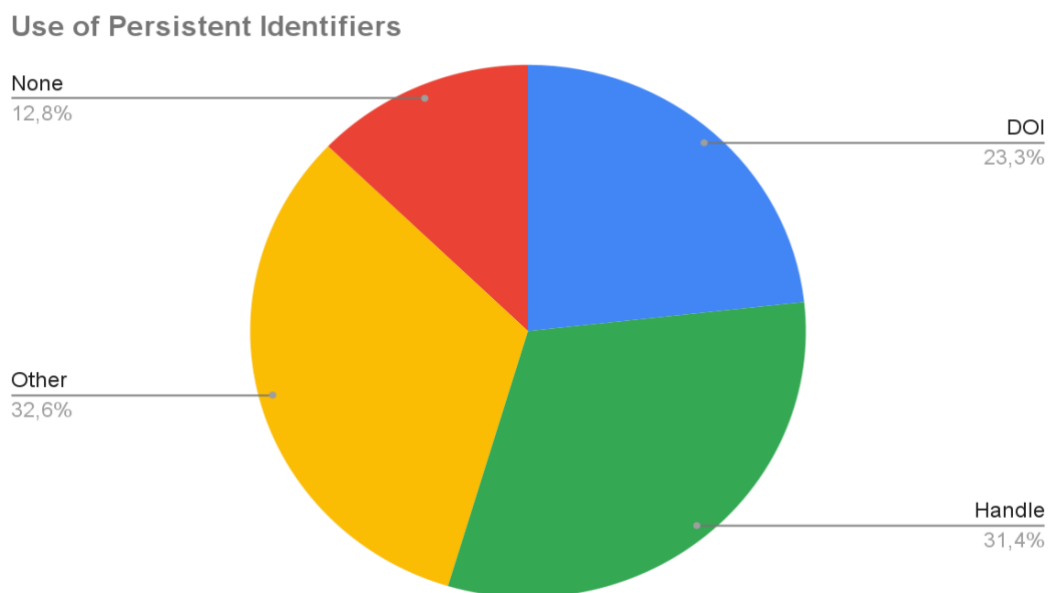
- Crossref: <https://github.com/CrossRef/rest-api-doc>
- DataCite: <https://support.datacite.org/v1.1/docs/api>
- mEDRA: <https://api.medra.org/>
- EIDR: <https://www.eidr.org/technical-documentation/>.

⁵² <https://api.crossref.org/swagger-ui/index.html> (accessed Sep 2021)

3.2 Analysis

The analysis of these repositories is represented below with graphs and textual explanations of the findings.

Figure 08: Use of Persistent Identifiers

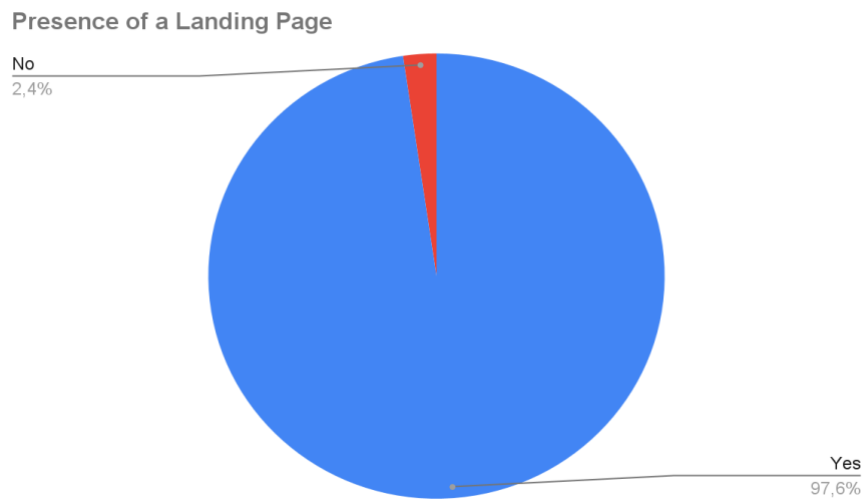


Persistent Identifiers are utterly vital for good data citation practices, as they ensure that researchers can continue to access the resource that is being cited. In 87.4% of the repositories the team examined, PIDs were used, distributed between DOI, Handle, and “other” forms of persistent identifiers, such as permalinks.

It should be noted that the analysis of these PIDs was rather uncritical, counting as “Others” all forms of permalinks, even in some cases the URLs do not seem to be permanent or well-structured.⁵³ Overall, though, it appears that the importance of having PIDs is broadly recognized by data repository creators and maintainers, even if there is still room for progress.

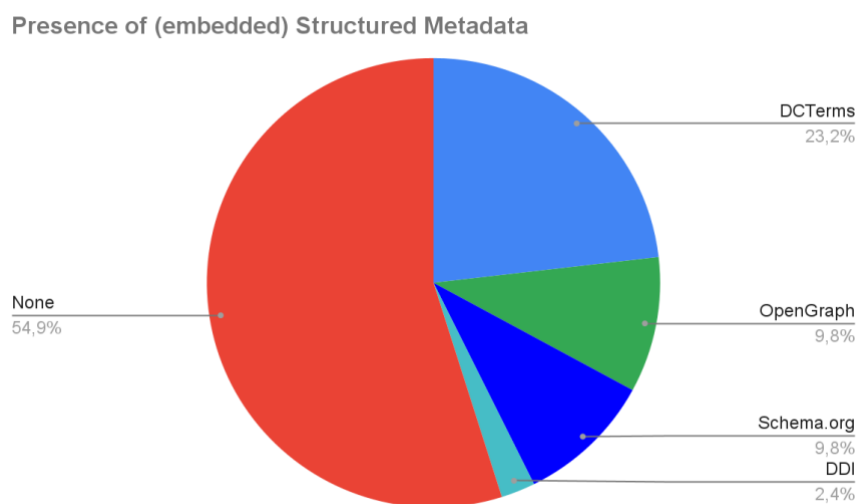
⁵³ e.g. <https://www.ids-mannheim.de/digspra/kl/projekte/korpora/archiv/mk>

Figure 09: Presence of a landing page



Landing pages appeared in 97.6% of the repositories the team examined, which is a recognition of their importance for communicating important metadata about the data they represent.

Figure 10: Presence of Structured Metadata

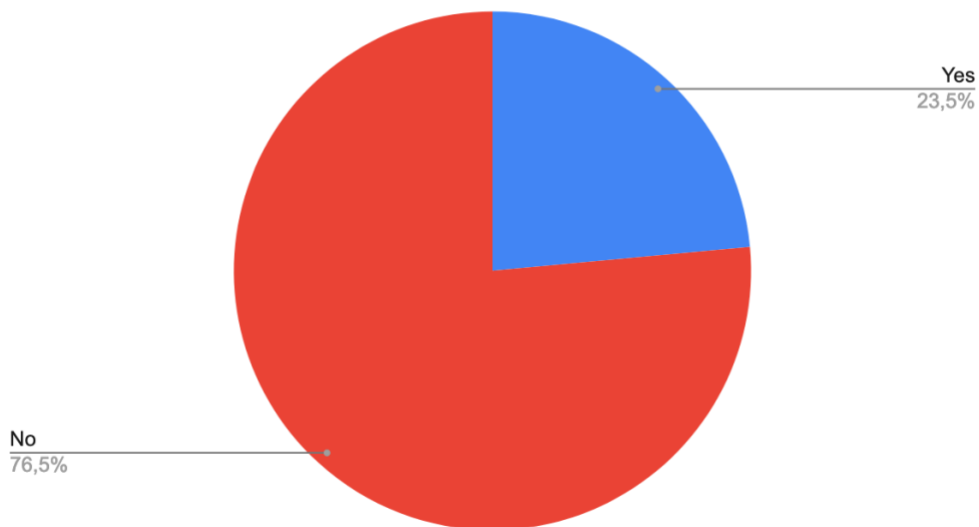


The presence of embedded, structured metadata in the source code of the webpage is an important and key way to make one's data more accessible by making it machine actionable. Using ontologies such as

DCTerms, OpenGraph, Schema.org or others, this encoded metadata allows a machine to automatically “read” the metadata associated with a given resource in a data repository. Our analysis shows that 45.1% of data repositories make use of this technology, with DCTerms representing nearly half of these repositories. However, it should be noted that when there was more than one encoded metadata ontology employed, the team chose the most-used ontology. Thus, the ontologies that the team discovered in our analysis include DCTerms, OpenGraph, Schema.org, DDI, Twitter, and CMDI even if the latter two are not represented on the graph. There is room for improvement, yet it appears that the importance of structured, embedded metadata is not ignored in the SSH community.

Figure 11: Presence of Versioning

Presence of Versioning



Only a quarter of repositories provided an access to different versions of hosted data. Additionally, even when if this possibility is available, the consistency of persistent identifiers regarding those different versions is generally far from perfect. It should be noted that most repositories only host one version of the data, which may explain the low number of versioning feature provided.

Figure 12: Ready to use "Cite As" feature

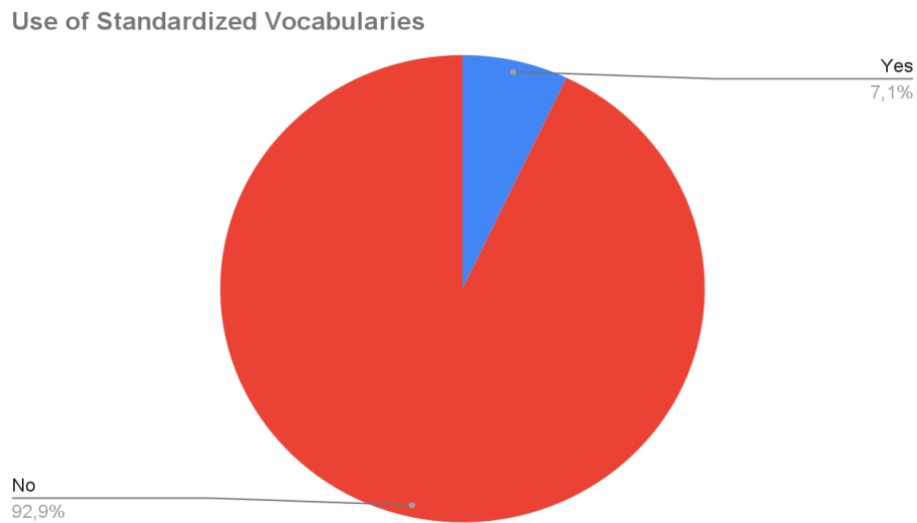
Ready to use "Cite As"



Nearly half of the repositories provided a ready to use "Cite As" option, which proposed one or many ways for researchers to cite the data contained in the repository. These "Cite As" properties included everything from simple textual renditions of a citation, a dialogue box in which one can choose from a selection of citation formats, an option to download the data in XML, or a combination of the above.

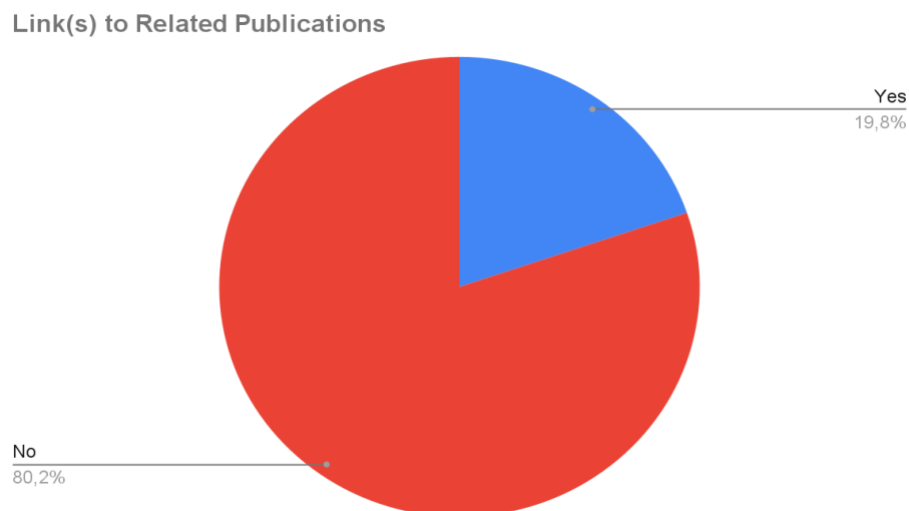
The newer data repositories had a ready to use "Cite As" function more often than older, less well-maintained data repositories.

Figure 13: Use of standardized vocabularies



Our analysis also checked for the use of standardized vocabularies. In principle, this meant that the team looked for the presence of an ORCID link for the data creators, which the team found only very rarely and also Creative Commons for licences which is more frequent. This is a potential area of improvement as the field moves forward, as a way to better increase links between existing resources (e.g. from Linked Open Data). It should be noted that some data repositories did semantically link authors within their own database, thus allowing a local version of the proposed data web, though these were rather rare cases and did not count towards the use of standardized vocabularies, as they were wholly internal.

Figure 14: Link(s) to related publications



The final area of analysis was the presence of link(s) to publications which used the data linked in the repositories. Nearly one in five data repositories took this important step, which allows researchers to easily discover the scientific productions related to a given data entry. However, these links towards publications were regrettably not always standardized and did not uniformly contain PIDs for the accompanying publication. Indeed, sometimes these publication links were merely a string rather than any attempt at semantic linking.

General overview of the results

Taken together, there is only one repository which fulfils all the criteria described previously. Most, a combined 62.3% of repositories, meet three or more of the criteria. Overall, there is room for improvement. A good target could be to have most repositories providing between four and five criteria, a standard which is met by only 40.3% of the repositories surveyed.

Figure 15: Number of Total Criteria Fulfilled

Number of Total Criteria Fulfilled

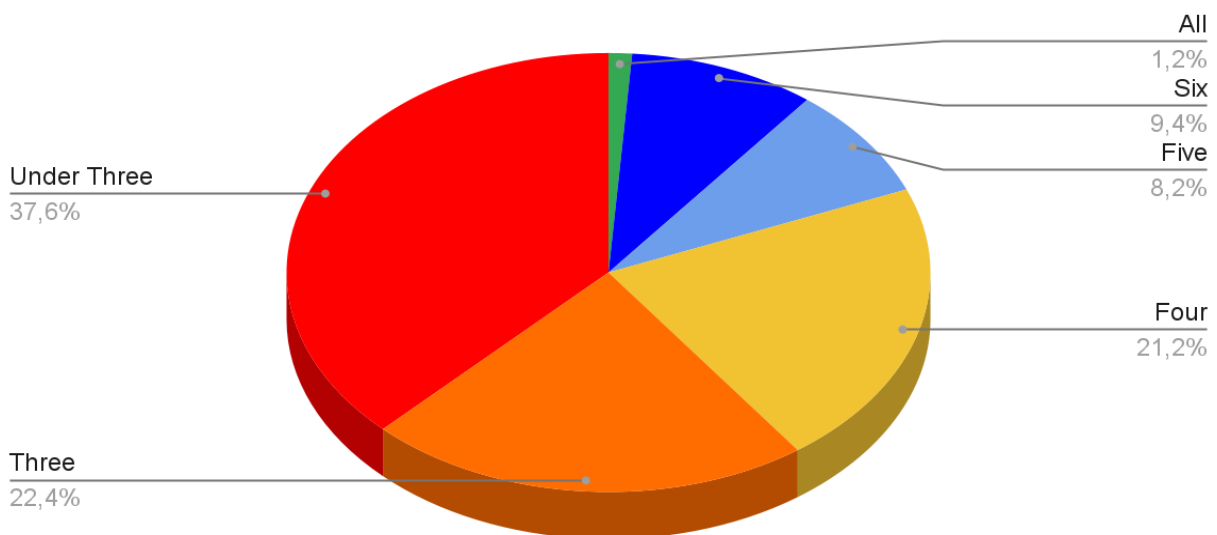
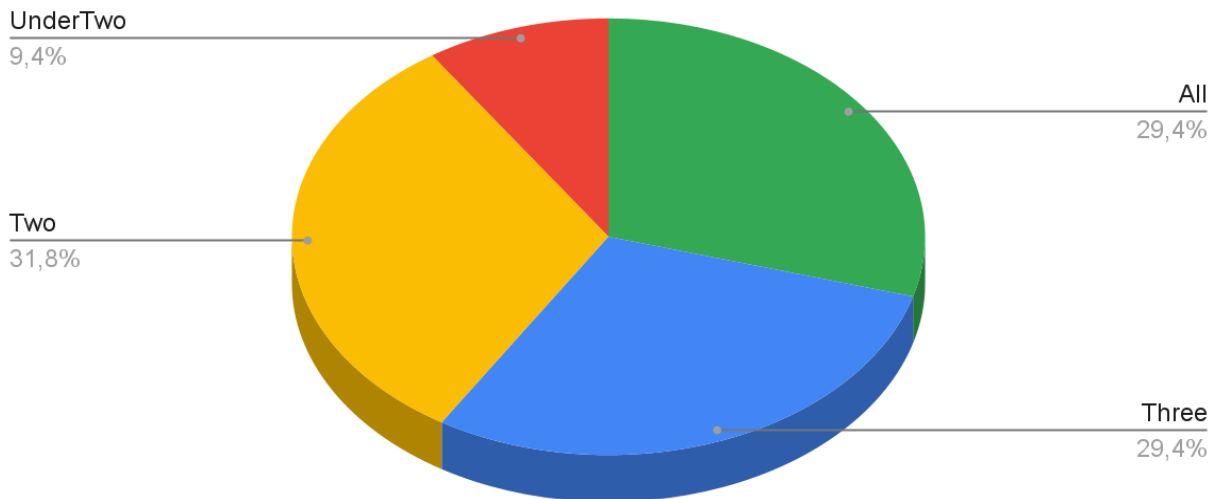


Figure 16: Number of (4) Main Criteria Fulfilled

Number of (4) Main Criteria Fulfilled



If the analysis is limited to the four main criteria mentioned in section 1.3 (i.e. PID, Landing Page, Structured Metadata, Cite As), the data are more encouraging as nearly a third of repositories meet the four criteria judged necessary to build robust data citations using these infrastructures. Nevertheless, there remains room for improvement as 40% meet fewer than 3 criteria.

4. Conclusion - Ways Forward in SSH Data Citations

Overall, the results of our data repository analysis are encouraging, even if they demonstrate that there are still areas of improvement. The near-ubiquitous presence of landing pages is an important building block in building a robust data citation environment in SSH even if they do not all contain information suitable for machine actionability. As well, the very high percentage of persistent identifiers bodes well for the future sustainability of these data citations, especially when coupled with the growing presence of ready to use “Cite As” feature. The detailed analysis of individual repositories carried out in Section 2 also confirms that repositories can adapt to the needs of their particular target audiences, be they

institutional repositories, or broader social science, humanities, or infrastructure-based projects. This explains the diversity of services provided by these repositories regarding data citation in general.

However, the team have noticed that the data produced are quite difficult to find on institutional web sites and are often totally absent. It's easy to find a description of research done (themes, teams, projects), events, funding received, publications etc. but rather more difficult to find research data. A possible improvement would be to encourage institutions to create a clear space or at least a link dedicated for research data on their websites. If one can consider that descriptions of research activities, events, grants, and publications reflect the priorities in these labs, it is possible to envisage funding agencies requiring these deposits to be clearly accessible, in much the same way as it was funding agencies that stimulated the increasing use of Data Management Plans. Additionally, even when you have a link to find hosted data, it is not always easy to retrieve relevant information; e.g. it's not clear how to make a simple query to find a corpus and sometimes the search engine is not really efficient.

The use of PIDs is very diverse. Some are not really precise (e.g. permalink or simple URLs). There is clearly a need for proper infrastructures to be able to cite data “durably” by maintaining PIDs and thus access to data. Furthermore, proper infrastructures could help ensure that metadata are machine readable to create actionable citation in order to enhance the dissemination of research products.

There are other areas of improvement for the future. Some repositories provide links to different formats of metadata in the landing pages, but nothing that is directly embedded into the page, and therefore explicitly machine actionable. Moreover, the Cite As functionality, despite an encouraging uptake of 49.4% among the linked surveys, is also subject to a great diversity ranging from a simple string to fully actionable dialogue boxes that allow the user to select from a variety of bibliographic norms and download formats in BibTex or XML. Indeed, while it is not sufficient to simply have a string that contains no dynamic data, data repository managers should also avoid only providing downloads to BibTex or XML with no other options to access the preferred citation. The best practice remains to multiply the options to fit the needs of a diverse group of researchers.

The usage of the citation viewer in the FAIR Citation Prototype was really a key to carry out this survey. As a side effect, by using this part of the prototype the team were able to identify possible improvements for the Citation Prototype that the team are going to implement before the end of the project as it is now more a “Proof Of Concept” than a running service:

- During this survey, the team used only the first type of metadata observed (e.g. schema.org) as the viewer searches sequentially per type. It could be interesting to check the presence of other metadata to enrich the description of the dataset
- Also concerning metadata enrichment, other sources of metadata should be considered, for instance sometimes some links to other formats are provided (e.g. DDI format) from the landing page and the team could try to identify these types of links

- Regarding the content of metadata, the team had a quick look at controlled vocabularies focussing mainly on authors with ORCID and licences with Creative Commons. A more systematic search for controlled vocabularies should be done (e.g. TADIRAH for the description, GeoNames for places, DOIs for publication etc.) with the prototype in order to propose better information and lay the groundwork for building a graph linking data, publications, places, authors etc. Additionally, if the description of the type of data is normalized, the team can link that data to different tools; the SSHOC prototype already offers links with the Switchboard⁵⁴ that is being further developed in the context of SSHOC
- Another interesting source to improve the quality of metadata would be to use “external sources”. For instance, enriched information provided by “registrars” like DataCite or Google Dataset (e.g. links to publications).

SSH are makes use of dynamic data, for instance data coming from Social Networks (e.g. Twitter).⁵⁵ This type of data raises new questions that will require adaptations of practices for their citation and the development of specific repositories.

All these underlying efforts to provide data citations, from infrastructures to good and normalized practices, pave the way for one of the next big steps in the evolution of data citation in the SSH: the creation of data papers. These data papers may hopefully lead to the creation of peer reviewed data journals: these elements are for the moment not really present in the world of SSH but attempts exist⁵⁶ and the team expects them to grow gradually. These technologies and practices will make it easier to implement links between data and publications. This will promote the dissemination of SSH research results by giving them greater visibility and also credit to the different stakeholders involved in the creation of research data.

⁵⁴ <https://sshopencloud.eu/sshoc-switchboard> (accessed Sept 2021)

⁵⁵ See, for instance, Breuer, Johannes, Borschewski, Kerrin, Bishop, Libby, Vávra, Martin, Štebe, Janez, Strapcova, Katarina, & Hegedűs, Péter. (2021). Archiving Social Media Data: A guide for archivists and researchers (1.2). Zenodo. <https://doi.org/10.5281/zenodo.5041072>

⁵⁶ Example from JODH journal <https://openhumanitiesdata.metajnl.com/articles/10.5334/johd.9/>

5. References

Publications

- Larrousse, N., Broeder, D., Brase, J., Concordia, C., & Kalaitzi, V. (2019). SSHOC D3.2 Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning. <https://doi.org/10.5281/ZENODO.4436736>
- CODATA-ICSTI Task Group on Data Citation Standards and Practices, (2013). Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. *Data Science Journal*, 12(0), CIDCR1–CIDCR75. <https://doi.org/10.2481/dsj.osom13-043>
- Andreassen, H. N., Berez-Kroeker, A. L., Collister, L., Conzett, P., Cox, C., Smedt, K. D., ... Research Data Alliance Linguistic Data Interest Group. (2019). Tromsø recommendations for citation of research data in linguistics (Version 1). Research Data Alliance. <https://doi.org/10.15497/RDA00040>
- National Research Council. 2012. For Attribution: Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop. Washington, DC: The National Academies Press. <https://doi.org/10.17226/13564>
- Carlo Maria Zwölf (VAMDC), Peter Wittenburg (RDA Europe), Zsuzsanna Szeredi (Vision & Values) GEDE RDA report (Pre-Print) https://github.com/GEDE-RDA-Europe/GEDE/blob/master/Citation/Report/GEDE_DC_report.docx
- Rauber, A., Gößwein, B., Zwölf, C., Schubert, C., Wörister, F., Duncan, J., Flicker, K., Zettsu, K., Meixner, K., McIntosh, L., Jenkyns, R., Pröll, S., Miksa, T., & Parsons, M. (2021). Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data. Zenodo. <https://doi.org/10.5281/ZENODO.4571615>
- Fenner, M., Crosas, M., Grethe, J.S. et al. A data citation roadmap for scholarly data repositories. *Sci Data* 6, 28 (2019). <https://doi.org/10.1038/s41597-019-0031-8>
- Rauber, Andreas, Asmi, Ari, van Uytvanck, Dieter, & Proell, Stefan. (2015, October 20). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). <http://doi.org/10.15497/RDA00016>
- Librarians' Competencies Profile for Research Data Management, 2016, https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf
- Akinici, B., & Bertone, A. (2017). Data Management Plan. Zenodo. <https://doi.org/10.5281/ZENODO.1240420>
- Breuer, Johannes, Borschewski, Kerrin, Bishop, Libby, Vávra, Martin, Štebe, Janez, Strapcova, Katarina, & Hegedűs, Péter. (2021). Archiving Social Media Data: A guide for archivists and researchers (1.2). Zenodo. <https://doi.org/10.5281/zenodo.5041072>
- Erzsébet Tóth-Czifra. (2020). 10 practical tips to fight against the culture of non-citation in the humanities. *DARIAH Open*. <https://dariahopen.hypotheses.org/747>

- Jennifer Edmond & Erzsébet Tóth-Czifra. (2018). Open Data for Humanists, A Pragmatic Guide. Zenodo. <http://doi.org/10.5281/zenodo.2657248>
- Colavizza, G. and Romanello, M., 2017. Annotated References in the Historiography on Venice: 19th–21st centuries. Journal of Open Humanities Data, 3, p.2. DOI: <http://doi.org/10.5334/johd.9>

Events Related to SSHOC task 3.4

- SSHOC / FREYA/ EOSC-HUB Realising EOSC Conference
FAIR Data-Citation for Social Sciences and Humanities
<https://www.eosc-hub.eu/events/realising-european-open-science-cloud/fair-data-citation-ssh>
- RDA Rich Metadata for annotation of citations contexts and data-citations contexts (BoF during 17th RDA Plenary) <https://www.rd-alliance.org/rich-metadata-annotation-citations-contexts-and-data-citations-contexts>
- SSHOC Round Tables of experts on Data Citation
<https://sshopencloud.eu/news/roundtable-experts-data-citation>
- SSHOC Workshop - Data Citation in Practice
<https://sshopencloud.eu/events/sshoc-workshop-data-citation-practice>

Other references

- FAIR Principles, <https://www.go-fair.org/fair-principles/>
- Data Citation Principles, <https://www.force11.org/datacitationprinciples>
- Force 11, Data Citation Principles, <https://www.force11.org/datacitationprinciples>
- Contributor Roles Taxonomy, NISO, <http://credit.niso.org/>
- Registry of Research Data Repositories, <https://www.re3data.org/>
- Citation Typing Ontology <https://sparontologies.github.io/cito/current/cito.html>
- Creative Commons, <https://creativecommons.org/share-your-work/>
- World Wide Web Consortium, Data On the Web Best Practices, <https://www.w3.org/TR/dwbp/>
- Open Data Commons, Legal Tools for Open Data, <https://opendatacommons.org/licenses/odbl/>
- OSSDIP: Open Source Secure Data Infrastructure and Processes: https://www.ifs.tuwien.ac.at/~andi/secure_data_infrastructure.html
- Digital Trace Data, https://sicss.io/2019/materials/day2-digital-trace-data/what-is-digital-trace-data/What_is_Digital_Trace_Data.html#
- Digital Preservation Risk Matrix, https://github.com/usnationalarchives/digital-preservation/tree/master/Digital_Preservation_Risk_Matrix

List of Figures

[Figure 1: Extraction of a recommendation from “Importance” principles which is applicable to a repository](#)

[Figure 2: Example of a citation string generated by a Dataverse platform](#)

[Figure 3: Metadata from CoCoon Landing Page expressed in various citation styles](#)

[Figure 4: The DARIAH-DE Repository Architecture](#)

[Figure 5: Citation extracted from the landing page of DOI 10.34847/nkl.7847b549](#)

[Figure 6: Sample of table used to check repositories](#)

[Figure 7: Server-driven HTTP Content Negotiation](#)

[Figure 8: Use of Persistent Identifiers](#)

[Figure 9: Presence of a landing page](#)

[Figure 10: Presence of Structured Metadata](#)

[Figure 11: Presence of Versioning](#)

[Figure 12: Ready to use “Cite As” feature](#)

[Figure 13: Use of standardized vocabularies](#)

[Figure 14: Link\(s\) to related publications](#)

[Figure 15: Number of Total Criteria Fulfilled](#)

[Figure 16: Number of \(4\) Main Criteria Fulfilled](#)

Appendix: List of repositories examined

Repository Name	URL
Finnish Social Science Data Archive	https://www.fsd.tuni.fi/fi/
Austrian Social Science Data Archive	https://aussda.at
Social Sciences and Humanities Data Archive	https://www.sodha.be/
Czech Social Science Data Archive	http://archiv.soc.cas.cz/
Danish National Archives	https://www.sa.dk
PROGEDO Research Infrastructure	http://www.progedo.fr http://quetelet.progedo.fr/
GESIS - Leibniz Institute for the Social Sciences	http://www.gesis.org/
Greek research infrastructure for the social sciences	http://sodanet.gr
Tárki Data Archive	http://www.tarki.hu/en/
Data Archiving and Networked Services	https://dans.knaw.nl/en
NSD - Norwegian Centre for Research Data	http://www.nsd.no
Portuguese Social Information Archive	http://www.apis.ics.ul.pt/
Slovak Archive of Social Data	http://sasd.sav.sk/sk/
Social Science Data Archives	https://www.adp.fdv.uni-lj.si/
Swedish National Data Service	https://snd.gu.se/en
Swiss Centre of Expertise in the Social Sciences	http://forscenter.ch/en/
UK Data Service	http://ukdataservice.ac.uk
Data Center Serbia for Social Sciences	https://datacentarserbia.com/en/
ASV Leipzig	https://centres.clarin.eu/centre/4
ACDH - A Resource Centre for the HumanitiEs	https://centres.clarin.eu/centre/45
Bayerisches Archiv für Sprachsignale	https://centres.clarin.eu/centre/5
Berlin-Brandenburg Academy of Sciences and Humanities	https://centres.clarin.eu/centre/6
Center of Estonian Language Resources (CELR-EKK)	https://centres.clarin.eu/centre/15
The CLARIN Centre at University of Copenhagen	https://centres.clarin.eu/centre/14

CLARIN-PL Language Technology Centre	https://centres.clarin.eu/centre/25
CLARINO Bergen Center	https://centres.clarin.eu/centre/29
CLARIN.SI Language Technology Centre	https://centres.clarin.eu/centre/30
CMU-TalkBank	https://centres.clarin.eu/centre/18
Eberhard Karls Universität Tübingen	https://centres.clarin.eu/centre/1
Hamburger Zentrum für Sprachkorpora	https://centres.clarin.eu/centre/9
Institut für Deutsche Sprache	https://centres.clarin.eu/centre/11
Institut für Maschinelle Sprachverarbeitung	https://centres.clarin.eu/centre/10
Instituut voor de Nederlandse Taal	https://centres.clarin.eu/centre/22
LINDAT/CLARIN	https://centres.clarin.eu/centre/3
MPI for Psycholinguistics	https://centres.clarin.eu/centre/24
PORTULAN CLARIN	https://centres.clarin.eu/centre/50
Språkbanken, The Swedish language bank	https://centres.clarin.eu/centre/37
The ILC4CLARIN Centre at the Institute for Computational Linguistics	https://centres.clarin.eu/centre/34
The Language Bank of Finland	https://centres.clarin.eu/centre/17
Universität des Saarlandes	https://centres.clarin.eu/centre/13
NAKALA	https://www.nakala.fr/
Geisteswissenschaftliches Asset Management System	http://gams.uni-graz.at
Digital Repository of Ireland	https://www.dri.ie/
Digital Repository of Scientific Institutes	http://rcin.org.pl
Online Digital Source and Annotation System	https://www.odsas.net/
Archive of the Italian Latinity of the Middle Ages	http://en.alim.unisi.it/
Culturaitalia	http://www.culturaitalia.it
Digital library of late-antique latin texts	http://www.digiliblt.unipmn.it/index.php
Digital Library Federation	http://fbc.pionier.net.pl/
Collection of documents (text, sheet music, audio, video, photo) about traditional and contemporary culture and society	https://repositorij.dief.eu/
Linguistic service, portal of many dictionaries	https://fran.si/

The open archive HAL	https://hal.archives-ouvertes.fr/
History of Slovenia	http://www.sistory.si/
IAH online library catalogue and database	http://library.foi.hr/lib/index.php?B=561
Digital Collections in the Cloud	https://locloudhosting.net/
A web museum to show and share videos, documentaries and studies related with cultural manifestations of Intangible Cultural Heritage (ICH).	http://www.memoriamedia.net/index.php/en
Digital Archives for Medieval Culture	http://www.mirabileweb.it/
Polish Literary Bibliography	https://pbl.ibl.waw.pl/
Portuguese Early Music Database	http://pemdatabse.eu/
The Institutional Repository of the Universidade Nova de Lisboa	https://run.unl.pt/
The TextGrid Repository is a digital preservation archive for human sciences research data	https://textgridrep.org/
Repository of Institute of Slovenian Ethnology	http://isn3.zrc-sazu.si/etnofolk/OAI-2.0/oai.php?verb=ListRecords&metadataPrefix=eef
The ISIDORE research platform	https://isidore.science/
Online repository of texts in the field of Slavic studies	https://ispan.waw.pl/ireteslaw/
DARIAH-DE Repository ingest tool	https://de.dariah.eu/en/publikator
Bibliographic database of world Slavic Linguistics publications	http://www.isybislaw.ispan.waw.pl/
Piattaforma Lessicografica Unica del Tesoro delle Origini	http://pluto.ovi.cnr.it/btv
Corpus testuale OVI	http://www.ovi.cnr.it/Il-Corpus-Testuale.html
Koninklijk Instituut voor het Kunstpatrimonium	www.kikirpa.be
Koninklijke Musea voor Kunst en Geschiedenis/Museés Royaux d'Art et d'Histoire	http://www.kmkg-mrah.be/fr/archives
Laboratoire de Recherche des Monuments Historiques	https://www.lrmh.fr/centre-de-ressources-synapse.aspx
Stiftung Preussischer Kulturbesitz - Rathgen-Forschungslabor - Staatlichen Museen zu Berlin	https://www.smb.museum/museen-und-einrichtungen/rathgen-

	forschungslabor/service/dienstleistungen.html
National Institute for Heritage	https://patrimoniu.ro/
Instituto del Patrimonio Cultural de España	https://ipce.culturaydeporte.gob.es/inicio.html
Basel Mission Archives	http://www.bmarchives.org/
National Gallery	https://www.nationalgallery.org.uk/research/research-centre/archive
Centro Nacional de Investigación sobre la Evolución Humana	https://www.cenieh.es/
Université de Lille Open Archive	https://lilloa.univ-lille.fr/
Foundation for Research and Technology - Hellas	https://www.ims.forth.gr
Vrije Universiteit Amsterdam	https://www.vu.nl/en/
Science for Life Laboratory	https://www.scilifelab.se/
University of Ljubljana, Faculty of Chemistry and Chemical Technology	https://www.uni-lj.si/academies_and_faculties/faculties/2013071111393229/
BioArCh	https://pure.york.ac.uk/portal/en/datasets/searchall.html?searchall=dataset
University College London	https://discovery.ucl.ac.uk/
Center for Socio-Political Data / Sciences Po	https://data.sciencespo.fr/dataverse/cdsp https://cdsp.sciences-po.fr/en/