Research and Innovation Action

Social Sciences & Humanities Open Cloud

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

Deliverable 3.2

Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning

Dissemination Level	PU
Due Date of Deliverable	30/06/19 (M6)
Actual Submission Date	15/11/19
Work Package	WP3 - Lifting Technologies and Services into the SSH Cloud
Task	T 3.4 Making Data Findable by being Citable
Type	Report
Approval Status	Approved by the EC - 03 November 2020
Version	V 1.0
Number of Pages	p.1-p.34

Abstract: This deliverable is a report on Data Citations In SSH. It pertains to Task 3.4 under the responsibility of CNRS. In addition to an inventory of Data Citation Practices, this deliverable includes recommendations for the SSHOC project in particular for WP7 activities (SSHOC Marketplace).

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

History

Version	Date	Reason	Revised by
0.0	10/05/2019	First Draft	Nicolas Larrousse
0.1	29/05/2019	Agreement on structure & References	Daan Broeder
0.2	13/06/2019	Intermediate version	Everyone on the editorial team
0.3	13/08/2019	Peer review	Coordinator
0.4	18/10/2019	Peer review comments addressed	Nicolas Larrousse
1.0	11/11/2019	Final Version after proofreading	Nicolas Larrousse

Author List

Organisation	Name	Contact Information
CNRS	Nicolas Larrousse	Nicolas.Larrousse@huma-num.fr
CLARIN-ERIC	Daan Broeder	daan.broeder@di.huc.knaw.nl
UGOE	Jan Brase	brase@sub.uni-goettingen.de
CNR	Cesare Concordia	cesare.concordia@isti.cnr.it
LIBER	Vasso Kalaitzi	Vasso.Kalaitzi@kb.nl

Contributor List

Organisation	Name	Contact Information
LIBER	Athina Papadopoulou	athina.papadopoulou@kb.nl
FSD	Henri Ala-Lahti	henri.ala-lahti@tuni.fi
University of Essex	Hervé L'Hours	herve@essex.ac.uk
SND	Birger Jerlehag	birger.jerlehag@snd.gu.se

Executive Summary

The SSHOC project aims to build the SSH (Social Science and Humanities) part of the EOSC (European Open Science Cloud). One of the main goals of the project is to ensure that SSH will be present in EOSC and that their specifics are taken into account.

In this regard, an important point is to be able to give high visibility to the research data used in Social Science and Humanities following FAIR data principles. This can be achieved by fostering "Data Citation" through providing a common mechanism to cite SSH data and build stronger links between data and publications. An expected side effect would be to enhance the reproducibility of SSH research, which is not very common nowadays.

In line with these goals, this report delivers an overview of existing data-citing mechanisms intended for citing data, that are currently used in different communities with a focus on the SSH. We then provide some guidelines describing what we think is relevant in the SSHOC context with respect to the technology to be implemented and also about the structuring of the content.

This report aims to prepare further work in common with other SSHOC tasks and Work Packages. Currently, we see the following links with other SSHOC tasks:

- SSHOC task 3.6 will generalise and integrate the CLARIN Language Resource Switchboard (Switchboard) for the infrastructure of the other SSHOC stakeholders. The Switchboard will also make citations human actionable. SSHOC task 5.2 works on a SSHOC version of the DataVerse repository system for the SSH. The citations function in SSHOC DataVerse should align with the recommendations put forward in this document.
- collaboration in engagement and training activities of WP6; it is essential to be able to give a solid overview of practices and recommendations in the SSHOC training and outreach effort. WP6 will coordinate targeted training in the Social Sciences and Humanities. Appropriate audiences will be targeted for a workshop and a webinar dedicated to "Data Citation principles and practice".
- WP7 is creating the future SSHOC Marketplace. This Marketplace will, of course, integrate SSH data sets and associated services. In this respect, a system that allows an extensive scientific description of data, strong identification of provenance and information that makes it possible to associate a relevant tool to a data set and make it actionable, would prove useful. This kind of information should be provided by Data Citation recommendations in cooperation with WP7 and other Work Packages.
- WP8, more specifically on certification aspects provided by task 8.2, "Trust and Quality Assurance".

Naturally, we should be involved in SSHOC activities whenever citation expertise is called on.

Abbreviations and Acronyms

ARK	Archival Resource Key
CDL	California Digital Library
CESSDA	Consortium of European Social Science Data Archives
CLARIAH	Common Lab Research Infrastructure for the Arts and Humanities
CLARIN	Common Language Resources and Technology Infrastructure
CNRS	Centre National de la Recherche Scientifique
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DASISH	DAta Service Infrastructure for the Social sciences and Humanities
DMP	Data Management Plan
DOI	Digital Object Identifier
EOSC	European Open Science Cloud
E-RIHS	European Research Infrastructure for Heritage Science
ESS	European Social Survey
EUDAT	EUropean DATa Consortium
FAIR	Findable Accessible Interoperable Reusable
Handle	Persistent Identifier System provided by CNRI
IIIF	International Image Interoperability Framework
LIBER	Ligue des Bibliothèques Européennes de Recherche – Association of European Research Libraries
OCLC	Online Computer Library Center
OPENAIRE	Open Access Infrastructure for Research in Europe
PID	Persistent Identifier
PURL	Persistent uniform resource locator
RDA	Research Data Alliance
URN	Uniform Resource Name
W3C	World Wide Web Consortium

Table of Contents

1.Introduction: Why data citation?

Among SSH communities, research data were largely ignored or considered as auxiliary by funders and even, to a certain extent, by researchers. Nowadays, principally for the purpose of reproducibility and transparency of the research process, there is a need to provide access to research data. In order to do so, it is necessary not only to preserve these data but also the associated information that would render them findable and reusable. This information should cover a broad scope from date of creation to provenance, so that the researcher can understand how to use the data and also the data history. Even if providing access is placed at the very end of the research cycle, the process of creating this information has to be considered from the very beginning of the project.

Forerunners in preserving and giving access to research data are traced back to astronomical observatories. Data in astronomy is generally expensive to produce and sometimes is simply impossible to reproduce at a different period of time considering the astronomical timeline. Notwithstanding, some data are historical and still relevant and can be considered as a heritage to be preserved.

Likewise, in the SSH a strong motivation is the fact that funding agencies realized that creating data for research purposes is extremely costly, in terms of financial and human resources. From the agencies point of view, this investment needs to be considered very seriously, which has led to the introduction of Data Management Plans (DMP) to be created at the very beginning of a project. In many cases, (e.g. in EU funded projects[1]) this procedure is mandatory and clear even at the stage of the request for funding. The DMP is now an important part of the evaluation of a project. As mentioned in B. Schmidt et al.[2], based on funder requirements the need to support researchers in creating and implementing data management plans has substantially grown in recent years. Several libraries have set up a service to support such needs, often in collaboration with other service units (e.g. research office, IT services, legal advisor, ethics committee). The development of such a service can even serve as a training ground for librarians and other institutional stakeholders (David &; Cross, 2015).

Other initiatives, such as the FAIR principles[3], were also essential in establishing the importance of data in the research landscape not to mention of "Open Science crusades".

As pointed out above, the notions of permanent and easy access to data are also crucial in this context. This means that we need to build proper infrastructures both to provide access and also to preserve data. The idea of infrastructures for SSH is quite recent but considering the huge increase in data generated for research purposes, they have successfully proved their effectiveness and usefulness.

If we have a standardised way to cite data that provides a robust basis for access, it will be easier to build and curate the relationship between scholarly literature and data. By doing so, it will be easier to achieve the

1 Open Research Data Pilot https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

2 Librarians' Competencies Profile for Research Data Management, 2016, https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf

3 The FAIR Guiding Principles for scientific data management and stewardship https://www.nature.com/articles/sdata201618

automatic discovery of articles linked to a specific dataset and vice versa. This will contribute to the improvement of the whole scientific ecosystem within SSH disciplines.

In brief, citing data is important for different reasons:

- Providing a way to reproduce research which will in turn enhance the quality and effectiveness of research
- Reusing data for different research purposes in other contexts
- Giving credit to the creator and the funder of the data
- Proving the usefulness of infrastructures
- Enhancing links between data and publications

The benefits of citing data have been developed in greater detail by the "CODATA-ICSTI Task Group on Data Citation Standards and Practices" in the paper "OUT OF CITE, OUT OF MIND"[4]. This paper also stressed the fact that there are also some real "cultural and institutional obstacles" to establishing data citation practices. Data are not necessarily considered as the noble part of the research process, especially within SSH communities. Moreover, even if researchers devote considerable effort to creating and processing data sets, it does not bring them any academic recognition.

The situation is evolving gradually with the emergence of Data Management Plans, which are one of the elements of the general "Data Stewardship wave" promoted by some groups like the GO-FAIR initiative[5].

While the citation of scholarly papers is already a well-established practice with a few internationally accepted standard practices or styles such as APA and MLA[6], citation of data is a relatively young and less standardised practice. However, although standardisation per se is relatively uncommon concerning this topic, some SSH communities have already developed their own ways of citing data, which have established themselves over the years in "communities of practice".

Nevertheless, Data Citation is a practice that has especially in recent years received much attention both in the context of research verification and reproduction and in the proper recognition of the efforts of the data scientists.

4 Task Group on Data Citation Standards and Practices of Cite, Out of Mind: The Current Sices, C.-l., 2013. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, 12, pp. CIDCR1–CIDCR7. DOI: http://doi.org/10.2481/dsj.OSOM13-043

5 GO FAIR Initiative https://www.go-fair.org/

6 MLA Formatting Guide https://owl.purdue.edu/owl/research_and_citation/mla_style/mla_style_introduction.html

2.Current Situation in SSH

In SSH communities, the situation is quite diverse. In Social Science, for instance, the Inter-university Consortium for Political and Social Research (ICPSR) proposed some recommendations to cite data hosted by the consortium in a common way[7]. In this field, therefore, the tradition of sharing data in Social Sciences has been quite well established for many years. In the field of Humanities, archaeologists used to consider data from archaeological excavations as an asset not be shared with others, to avoid competition, even though they cited, or rather referred to, this very set of data. This attitude is not unique: researchers in linguistics were not always eager to share their data with others. However, the situation is improving, as data creation and sharing is increasingly considered as part of the work.

Nevertheless, there is a strong need to be able to cite data and possibly individual parts of the data: e.g. researchers dealing with oral recordings need to cite part of a recording which can be reduced to a sentence in its context. In a completely different field, art historians explore data citation of artworks, for instance by referring to a part of an image through the use of IIIF technology. An example from the field of classical philology is the Canonical Text Services (CTS) that use an absolute URN:NBN based identifier system[8] to refer to text fragments.

Therefore, compared to other research domains, we can observe that there is a great variability in needs for SSH communities. Moreover, no strong tradition of data-sharing existed in Humanities communities, and as a result of this data citation is something relatively new in this area. The situation is rather different in Social Sciences where there is a fairly long tradition of data-sharing as well as data citation. It should also be noted that the notion of infrastructure for SSH is a relatively recent concept, especially compared to the duration of the research in the humanities.

As a result, in the SSH there has not been a specific common approach to data citation before. If there are common views, these are probably caused by the relative proximity of some of the SSH subdomains to the field of library science, the common participation in wider discipline-agnostic approaches as discussed and promoted by some Research Data Alliance[9] (RDA) working groups , and with reference to parts of the data citation workflow, by previous SSH collaboration projects such as the DASISH project[10] with respect to recommendations for the use of PIDs for data[11].

Within the SSH there are a few specific data-types, which could be shown to be invaluable for research, if a standardised practice for their citation were devised. The most important data-type is Social-Media data, which is becoming an increasingly important type of research data within the DARIAH, CLARIN and other research

7 ICPSR recommendation https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/citations.html
8 CTS URN http://www.homermultitext.org/hmt-docs/cite/cts-urn-overview.html
9 RDA https://rd-alliance.org
10 DASISH https://dasish.eu
11 DASISH D 5.1 Trust and PID Services Federation Report https://dasish.eu/publications/projectreports/DASISH-D5_1_BS_version141106.pdf

infrastructures: an illustration of this is the UK Web Archives that have been archiving[12] social media posts regarding the EU referendum among other web-based content.

The topic of dynamic data collections is a recurring one in the Humanities. As already mentioned, social media data are of increasing importance. However, some cases, for example, large newspaper article collections, where the availability of a specific article cannot be guaranteed for legal reasons, should be considered a dynamic collection as the content may vary over time. A metadata query defining a subset of articles will not return consistent results. This has led to the need to distinguish the concept of "intentional collections", which are dependent on a specific intention or query e.g. 'retrieve all articles from 10-10-2019', from extensional ones where the content is static.

Often dynamic data description and reference challenges are solved by using a strategy of creating regular "snap-shots" of data. These different versions may each warrant a separate metadata description and a separate PID. Decisions about this are considered to be a part of 'versioning policy'' e.g. the decision to issue separate PIDs and metadata for the different versions. Different organisations can choose different policies. In the CLARIN infrastructure, the CLARIN centres can each set their own policy, but are required to be explicit about their choice.

Web resources can be cited using MLA or APA format. Many of the recommendations made on how to cite a WikiPedia entry that can be reused in the context of data citation. Additionally, for resources that are no longer be directly accessible, the Internet Archive provides some information about the correct way to cite a URL from the Wayback Machine[13] in MLA format. However, this possibility should only be considered as a last resort for data citation as it will not generally provide metadata. . As for web archiving done at a national level, for instance in France the archiving by the French National Library, a URL is quite difficult to cite as it is only accessible internally.

Publishing Social Sciences and Humanities Datasets

Lawrence et al.[14] defined the act of 'publishing data' as: "to make data as permanently available as possible on the Internet."Published data should have been validated, to guarantee a reliable format and content, and should be easily accessible. Callaghan et al.[15] started from this definition and argued that the formal publication of data is more than just posting a dataset on a website, it should include:

- checks on the dataset of either technical or content-based nature
- description of the dataset using a set of metadata
- data persistence functionalities
- a platform for the dataset to be found, evaluated and possibly reused

12 UK Archive EU referendum https://www.webarchive.org.uk/cy/ukwa/collection/649

13 Using The Wayback Machine https://help.archive.org/hc/en-us/articles/360004651732-Using-The-Wayback-Machine

14 Lawrence, B., Jones, C., Matthews, B., Pepler, S., Callaghan, S.: Citation and peer review of data: moving toward formal data publication. Int. J. Digital Curation (2011). doi: 10.2218/ijdc.v6i2.205

15 Callaghan, S., Murphy, F., Tedds, J., Allan, R., Kunze, J., Lawrence, R., Mayernik, M.S., Whyte , A.: Processes and procedures for data publication: a case study in the geosciences. Int. J. Digital Curation 8(1) (2013). doi: 10.2218/ijdc.v8i1.253

Several software frameworks, called online data repositories or publication platforms, have been implemented to provide the publication functionalities. To cite a few: Dryad (http://datadryad.org), figshare (https://figshare.com), Dataverse (https://dataverse.harvard.edu), Mendeley Data (https://data.mendeley.com), Zenodo (https://zenodo.org) etc.

Publication platforms provide a number of advanced functionalities to manage data. For instance some of them support versioning to manage dynamic datasets (e.g. Dataverse[16] provides UI facilities to allow users to easily switch between different versions of a dataset and read details about metadata fields and files that were either added or edited in various versions), they can provide facilities for automatic metadata extraction (e.g. Zenodo[17] provides a module capable of extracting metadata from a variety of research data formats), and they may also provide software services to discover citation for a dataset (e.g. ScholeXplorer[18], developed in the OpenAIRE project).

Based on the available literature we find the following key points that should be considered in data publishing, in order to support a good policy for data citation:

- Use standard persistent identifiers to identify data.
- Consider that the digital object representing a non-digital data is distinct from its source and needs to be documented and cited in its own right.[19] The TEI Guidelines[20] explicitly require metadata for both the electronic object itself, and the data described by the electronic object.
- Adopt an efficient and reliable policy for dynamic data. If there are multiple versions of specific data, the metadata should describe their relation to other versions, and it should be possible to identify and cite a specific instance of the data. The RDA Dynamic Citation Working group has released a specification for citing data generated dynamically[21].
- Define and implement a clear copyright policy for data.
- Disclosure risk in data[22]. As preserving privacy and confidentiality in research data may be crucial in some social science datasets, citation of confidential data must be done in the context of legal agreement between the user and the publisher. A public version of a dataset may be required that coexists with the restricted-use version. These versions need to be distinguished and this may have implications for data citation.

A publisher can enable users to access datasets in several ways: by downloading them, by executing queries on them, by accessing them via specific APIs etc. A description of the access methods provided would be very useful, for instance publishing specification for the query language adopted and documenting APIs.

16 Data Verse Dataset management http://guides.dataverse.org/en/latest/user/dataset-management.html
17 Zenodo Meta Data Extraction https://gist.github.com/xiaom/106b9d111726cc99017b7efe7aead01c
18 OpenAire ScholeXplorer http://scholexplorer.openaire.eu/index.html#/
19 Michael Sperberg-McQueen: Data Citation in the Humanities: What's the Problem?, https://www.nap.edu/read/13564/chapter/10
20 TEI Guidelines https://tei-c.org/guidelines/
21 RDA Data Citation of Evolving Data https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf
22Mary Vardigan: Data Citation for the Social Sciences, https://www.nap.edu/read/13564/chapter/9

In general, data publishing means not simply publishing a dataset, but also publishing the related data article. Data articles (or data papers) describe the data, providing information on the methodologies followed to collect it, the format of the dataset, the license, the facilities provided to download data etc. The data themselves are not part of the article, instead they are linked to the article using persistent, unique links.

More specifically, we can find a list of 35 best practices was established by the W3C about publishing data on the web[23]. Clearly, these recommended best practises are closely linked to web technologies which is only one part, albeit an important one, in the process of publication of research data.

Even though not all these recommendations are usable in practice for research data management, when not directly applicable, they are at the least valuable sources of inspiration.

For instance, the best practice for Data Provenance (Best Practice 5) must be taken into account as use of W3C's Provenance Ontology should be considered in order not to "reinvent the wheel". On the other hand, "Provide data quality information" (Best Practice 6) is not really suitable for SSH data. Best practice 1, which proposes a possible integration of metadata in a web page by using web standards such as RDFA or JSON-LD, is a good starting point for structuring and implementing what we call a "landing page" later in this document.

Generally speaking, the proposed structure for the description of best practices: "Machine readable, Human readable, How to test" is a good role model.

We can also learn and build on the recommendation to "Enrich data by generating new data" part (Best Practice 31) which leads to what we call "semantic annotation" in that task: "data categorization, disambiguation, entity recognition, sentiment analysis and topification" are topics to be considered.

Data Citation for SSH Datasets

Social Sciences and Humanities (SSH) cover a large and heterogeneous set of disciplines. The datasets used in SSH are also heterogeneous and can vary from text documents (e.g. literature, survey data), to multimedia files (e.g. interviews or movies) to structured datasets (e.g. digitized census data). For instance, an archaeological dataset may consist of excavation photos, maps of the site, databases on the artefacts and a report.

On a large scale, communities or infrastructures promote best practices, standards and infrastructure to support the description and deposit of such resources. The DARIAH Research Infrastructure (RI) recommends the use of persistent Research Identifiers generated by a free, public and open service such as IdHals (https://doc.archives-ouvertes.fr/en/author-identifier-idhal-and-cv/) or ORCID (http://www.orcid.org/).

The CLARIN RI endorse the Force11 Data Citation principles [FN24] and CLARIN also requires the use of Handles as a PID system. To support the availability of Handle PIDs CLARIN joined ePIC, the European Persistent Identifier Consortium[24] that provides Handle PIDs for the research world.

23 Data On the Web Best Practices, https://www.w3.org/TR/dwbp/

24 European Persistent Identifier Consortium https://www.pidconsortium.eu

The social sciences have a long tradition of data citation. For instance, the Inter-university Consortium for Political and Social Research (ICPSR), already mentioned, has provided a citation standard for its data since the 1960s (Sieber and Trumbo[25], 1995) and the International Association of Social Science Information Services and Technology (IASSIST[26]) has done work on the bibliographic references of data files from early on (see Dodd[27], 1979 for example). The European social science data archives have followed the data citation guidelines of the field when deciding on their citation models.

Although the exact citation formats vary between organisations and individual archives, the basic elements, such as the author, title, date or year, and a persistent identifier are typically included. ICPSR suggests including information about the author, title, distributor, date, version and persistent identifier in the citation, while CESSDA[28] recommends including as minimal elements the creator, publication year, title, publisher and identifier. Individual CESSDA archives have their own distinct but similar citation models. A downside of the absence of a single standard or format is that the citations tend to be more free-form text strings and thus not readily machine-actionable. The CESSDA Persistent Identifier Policy[29] lists 6 main principles for the use of Persistent Identifiers (PID) across CESSDA Service Providers and contains guidelines to implement these principles.

With respect to the actual practice of citations we note that although practices vary in the SS, the use of DOIs is mostly accepted in the larger organisations. Also, there seems to be a need for acknowledging the role of all the funder and otherwise involved organisations, leading to sometimes very complex citations e.g. citation requirements for SHARE data sets[30] can lead to several paragraphs of text for a single citation. Other SS organisations, however, have simpler requirements, not using DOIs but rather URI links to institutional repositories.

Common features are that the identifiers always point to landing pages itemizing the different data-set components, available data formats and citation requirements, and that there is, unsurprisingly, a proper focus on version and publishing date as repeat surveys form a large part of the SS data.

The Data Archiving and Networking Services (DANS) adopt the citation guidelines recommended by the DataCite and FORCE11, each dataset has a unique DOI, and new datasets are automatically assigned a DOI. A noteworthy phenomenon is the appearance of peer reviewed data journals in SSH such as the "Research Data Journal for Humanities and Social Sciences" founded by DANS and published by Brill (see

25 Sieber, Joan E. & Bruce E. Trumbo (1995). "(Not) giving credit where credit is due: Citation of data sets". Science and Engineering Ethics 1: 1, 11–20.

26 IASSIST Data Citation Resources https://iassistdata.org/community/data-citation-ig/data-citation-resources

27 Dodd, Sue A. (1979). "Bibliographic References for Numeric Social Science Data Files: Suggested Guidelines". Journal of the American Society for Information Science 30: 2, 77–82.

28 CESSDA Citing your data https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Citing-your-data; CESSDA Access Use and Cite your data, https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/Access-use-and-cite-data

29 CESSDA ERIC Persistent Identifier Policy http://multiweb.gesis.org/csaw/#!Detail/cessda-eric/0047

30 SHARE Citation Requirements http://www.share-project.org/data-access/citation-requirements.html

https://brill.com/view/journals/rdj/rdj-overview.xml?lang=en). It has to be noted that some other journals, such as Cybergeo[31] in France, have opened up to data but more in an experimental way.

As another example, although mostly intended for cultural objects, Europeana has set up a specific task force[32] whose purpose is to identify a reliable and sustainable citation mechanism, in order to cite resources from Europeana in publications in general. Some communities such as archaeologists have organized themselves to cite data[33] in particular in the context of Cultural Heritage.

It can be seen that the interest in "data citation" at large, even if it is relatively recent, is broadly shared amongst different institutions and organizations and not only by infrastructures,

3. Landscape of existing technologies

3.1. Persistent Identifiers

Nowadays, most research data are now expressed in digital form and accessed via the Internet using URLs. However, a URI is difficult to maintain over time. For example, this URI, "http://my-lab.fr/my-project/my-corpus/my-data", depends on the domain name used (i.e. my-lab.fr), itself, in this case associated with an organization name that may disappear. Similarly, the rest of the URL (i.e. my-project/my-corpus/my-data) is semantically significant but will hardly survive a future reorganization or the introduction of different access protocols than HTTP.

W3C has insisted that nevertheless a URI should be used for identifying resources on the internet and advocated the use of so-called cool-URIs[34] emphasizing the use of stable domain names, but not every organization can provide such a stable domain name. Thus, the use of Persistent Identifiers (PIDs) for identifying resources and whose string value is independent from the resource's URL, has gotten broad acceptance. PIDs should come with a resolving system that can resolve the identifier to a URL. Note that the desire for a reliable resolver system means that we rule out the URN:NBN type of persistent identifiers, which although are a system well suited to identify resources, lack such a resolver system.

Different brands of PIDs often exist with their own specific resolver technology to translate PIDs into the resource URL, to make it as transparent as possible to the user. Each of these PID systems has advantages and disadvantages.

Some of the most commonly one used in our communities are:

31 "Presentation of Data Papers", Cybergeo: European Journal of Geography [En ligne], Data papers, http://journals.openedition.org/cybergeo/28922

32 Europeana Resource Citation task force https://pro.europeana.eu/project/resource-citation-object-identity-standardization

33 Marwick, B. and S. E. Pilaar Birch (2018) A Standard for the Scholarly Citation of Archaeological Data. *Advances in Archaeological Practice* https://doi.org/10.1017/aap.2018.3

34 Cool URIs for the Semantic Web http://www.w3.org/TR/2008/NOTE-cooluris-20081203/

- PURL (Persistent URL)

  A system implemented by OCLC (https://www.oclc.org/) that allows you to have your own resolver or use the one provided by OCLC. However, it is a simple tool for matching an identifier with a URL;

- Handle

  A proven and relatively inexpensive technology based on a system of delegating the creation and management of identifiers to a system of distributed service providers, but which is dependent on a technology that must be maintained; Handles also allow the use of a system of additional identifiers (part-identifiers) that can be used to identify specific parts or specific versions of resources[35];

- DOI (Digital Object Identifier)

  The system is based on the same "Handle" technology described above but is implemented by a consortium that has specific policies for different aspects of DOI use that guarantee its sustainability. On the other hand, membership of the consortium is subject to a fee. Other complementary services are offered, in particular the management of specific metadata that are often used by scientific journals;

- ARK (Archival Resource Key)

  ARK identifiers are proposed by the CDL (California Digital Library) and offer extended functionalities compared to a "Handle" type system, in particular the notion of "qualifier" which allows a certain flexibility. For example, it is possible to refer to several versions of the same document, each of which thus has a unique identifier. Similarly, the identifier can also be made independent of the resolution device. Implementing these richer functionalities requires specific skill, however.

What strategies should be used to choose an identification system? As can be seen, the choice is not simple but will depend on the level of functionality required: from a simple URL abstraction to a real documentation system. It is also necessary to take into account the resources at your disposal: the implementation of these technologies and the management of the updating of identifiers can be costly in terms of human resources. Finally, it should be noted that a common practice is the use of several identification systems, which makes it possible to combine their functionalities.

Some publication and data citation standards require the use of a specific PID brand such as the use of DOIs by DataCite. This led then to alternative recommendations to use DataCite.

An important feature of some PID systems is their ability that it can also be used to identify 'related' objects of the primary one. Such related objects can be for instance "parts-of", other versions of, or other representations of the original. This functionality is offered by the DOI and ARK services.. Note that for solutions which propose different serialisation formats, HTTP content negotiation is often the preferred way.

Many different communities have now formulated policies and recommendations on how to use PIDs in their research workflow. In 2017 the GEDE[36] group within RDA produced a report entitled "Consolidated

[35] EPIC API Part Identifier https://doc.pidconsortium.eu/guides/api-partial/
[36] RDA GEDE Webworkshop https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda

Assertions"[37] which gives a good overview of many of these policies, including those of discipline agnostic initiatives. From the SSH, the positions of CLARIN[38] and CESSDA[39] are well represented. For our purposes, this summary of SSH viewpoints indicates that Handle and DOI appear to be the preferred PID systems. When other identifier systems are also needed (e.g. URN:NBN), the Handle syntax should preferably be based on the same pattern (e.g. from CLARIN).

An important general statement in this report with respect to the value of using Handles and DOIs is that Handles are eligible for use for data at all stages in the Data Life Cycle (DLC) while DOIs should be used for published collections, after quality checks and additional metadata have been made available. So there seems to be a preference for DOIs (if available) for use in citations. Although there is of course no guarantee that the quality checks in question have been executed by the data provider: this issue can be solved by developing specific certifications or using existing ones

The role of PIDs in data citation can be considered to be threefold:

- providing reliable access to the data i.e. a reliable resolving mechanism
- referring to parts of larger objects or collections
- referring to different versions as important.

3.2. Metadata

Persistent identifiers represent the first step in being able to cite data but, considering what has been said above, SSH data are unusable without at least some contextual explanations, usually in the form of descriptive metadata.

It is important to note that the term metadata itself can sometimes be ambiguous because its meaning differs from one category of users to another: researchers, especially in SSH, considers metadata as a scientific work, computer scientists as a technical information, other communities as annotation etc. One example among others is the notion of "creator" which can mean many different things and represent different kinds of realities.

There is considerable diversity in the way in which researchers describe data. This diversity relates to both the content of the description and its structure. The most commonly used structures (or schema) are of course the two flavors of Dublin-Core (i.e. Dc and DcTerms) but there are also more specialized schemas: for instance, DDI[40] for Social Sciences, OLAC[41]for oral resources and even TEI[42] or at least TEI header for some text resources.

[37] RDA Persistent identifiers: Consolidated assertions https://www.rd-alliance.org/group/data-fabric-ig/outcomes/persistent-identifiers-consolidated-assertions
[38] CLARIN PID policy from https://www.clarin.eu/node/3965
[39] General CESSDA documentation is mentioned as source
[40] DDI Alliance http://www.ddialliance.org/
[41] OLAC http://www.language-archives.org/
[42] TEI Consortium https://tei-c.org/

The 'citation metadata'[43] can be a subset, superset, or identical to the descriptive metadata available in organisational repository systems. Usually however it is a subset targeted at quick inspection and classification, not so very different from citation metadata for publications. If we are looking for a common citation format for the SSH, part of the challenge is to define a set of citation metadata that could be suitable for all SSH disciplines.

The most popular citation guidelines (APA, MLA[44]) for publications were not developed for citation of datasets and therefore they recommend the rules for citing a web document, so adding a URI and data of access to the regular citation metadata Title, Author, Publisher, etc.

Pew Hispanic Center. *Changing channels and crisscrossing cultures: A survey of Latinos on the news media.* (Data file and code book). Washington, DC: Pew Research Center, 2004. Web. 19 Sep 2011. <http://pewhispanic.org/datasets/signup.php?DatasetID=5>

The ISO 690 citation style extension for electronic on-line resources is not very different

NASA, 2014, *NASA Event Reflects on Accomplishments of Mars Rover* [online]. [video]. 2014. [Accessed 11 October 2014]. Available from: https://www.youtube.com/watch?v=CpS919WF--8

Other style recommendations were from the start more specifically targeted at data citation: they take into account the fact that requirements differ among disciplines and therefore tend to specify core elements. Datacite recommendations are:

Creator (PublicationYear). Title. Publisher. Identifier

or, a more elaborate format:

Creator (PublicationYear). Title. Version. Publisher. ResourceType. Identifier

In general, there has been considerable work on providing guiding principles rather than exact formats for citation metadata. e.g. The Joint Declaration of Data Citation Principles[45] and the FAIR Guiding Principles[46] .

In addition to that, it is necessary to take into account the more "technical" needs regarding "data provenance", "versioning" and other specific information. With respect to the interpretation of metadata and citations by software agents, in particular all the information needs to be well structured (see also the chapter on "actionability").

[43] In the context of this document, citation metadata is the information that is embedded together with the data-set identifier in a publication. However, we have also seen this term also used in connection with the information available from the landing page for a data-set

[44] MLA handbook for writers of research papers, New York: Modern Language Association of America, 2009, 7th ed.

[45] Data Citation Synthesis Group (2014). Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11 https://www.force11.org/group/joint-declaration-data-citation-principles-final

[46] Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18

To be usable both by humans and by machines, it is necessary to have other information. Here are some examples:

For humans:

- a comprehensive description of the data set
- the context of creation, what was the original purpose
- the choices made during the creation process, such as restriction to a given population
- etc.

For machines:

- a common set of metadata structured in the same way
- use of standard descriptions for location, subjects, time etc.
- a mechanism to prove the provenance
- etc.

Furthermore, you may want to have additional information created by users or by machines by automatic classification that we can call "annotations". This latter type of information is useful to liaise data with published articles, other work by the same author, DbPedia entries etc.

The Metadata Working Group from DataCite published a schema for metadata to describe data sets in a common way[47] with three levels of obligation: Mandatory, Recommended or Optional. These metadata can be used to generate citations citation with different tools such as "DOI Citation Formatter"[48]. For the DOI "10.4000/books.pur.51737" published by OpenEdition, one obtains the following this citation:

```
Taillandier-Guittard, I. (Ed.). (2015). Métaphore et musique. Presses universitaires de Rennes. https://doi.org/10.4000/books.pur.51737
```

Many repository systems (Dspace, CKAN, Hydra, …) offer DataCite type citation creation either natively or via an extension. DataCite provides a list of such systems[49].

DataVerse also proposes recommendations for citing data[50]. The citation has is composed of five human readable components: the author(s), title, year, data repository (or distributor), and version number. Two other components are machine-readable: an independent identifier and a fingerprint.

```
Gary King; Langche Zeng, 2006, "Replication data for: When Can History be Our Guide? The Pitfalls of Counterfactual Inference", Harvard Dataverse, V2, http://hdl.handle.net/1902.1/DXRXCFAWPKUNF:3:DaYlT6QSX9r0D50ye+tXpA==
```

47 DataCite Metadata Schema https://schema.datacite.org/ and http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf

48 DOI Citation Formatter https://citation.crosscite.org/

49 DataCite Repository Software Integrations https://support.datacite.org/docs/repository-software-integrations

50 Data Citation in DataVerse https://dataverse.org/best-practices/data-citation

The Mendeley Data[51] repository, from Elsevier, proposes a simple way to cite a data set on the landing page of a data set. For example for the data set identified by the following DOI "10.17632/b7pfsj5kxk.1".

```
Jiang, Xinjian; Long, Tenghai (2018), "Production of supra-regular spatial sequences by macaque monkeys", Mendeley Data, v1 http://dx.doi.org/10.17632/b7pfsj5kxk.1
```

Alternatively, they also provide an access to metadata through an API[52]

3.3. Citation and data discoverability

In discussions about the use of PIDs and metadata, the concept and use of landing-pages crops up. In the context of the present document a landing page is a human readable web-page that contains both information on a resource (data-set or publication) and a link or information (PID, URI, email) for obtaining access to the resource. The content of the landing page may be (covertly) structured so that it can also be machine actionable. The landing page content is usually descriptive metadata for the resource formatted for human consumption. In many cases a landing page is the only route available to access the resource. Landing pages are often identified by PIDs and the stable anchors for resources e.g. Datacite, many CESSDA and CLARIN centers and DARIAH repositories.

In these cases, both the metadata and the PID function as a proxy for the resource itself. It can be important to keep the PID and/or metadata/landing page alive, even if the resource itself is lost or has been explicitly deleted. That way a user that is guided by a PID to the landing page can discover that the required resource was deleted for some specific reason instead of the PID being an administrative mistake e.g. by a "404 Not found" error.

In recent years, triggered by the relatively new interest in data within SSH communities, the need to link data and publications arose. The most important point, as already mentioned, was to ensure that data were created and treated in a consistent way to certify their quality. This should lead, at the end of the cycle, to allowing for the reproducibility of the process or at least to verifying scientific hypotheses.

For example, the French economist Thomas Piketty, has put the data used in his book "Capital in the Twenty-First Century" online so some critics were able to check the data consistency and start a controversy (see Wikipedia entry https://en.wikipedia.org/wiki/Capital_in_the_Twenty-First_Century).

Another benefit would be to be able to measure the value of a dataset by counting the number of citations produced. Thus, the creation of the dataset could be taken into account for the advancement of the researcher's career. In this context, we can mention the publication of peer reviewed data journals in SSH such as the "Research Data Journal for the Humanities and Social Sciences"[53] founded by DANS and published by Brill.

51 Mendeley Data https://data.mendeley.com

52 Mendeley API https://dev.mendeley.com/

53 Brill Research Data Journal https://brill.com/view/journals/rdj/rdj-overview.xml?lang=en

In order to implement this, the very first step is to define a common way to identify and designate a data set, or at least to build a matching mechanism between different systems. As a first approach, we can learn from the example of what is done for alignments in Semantic Web.

Once this prerequisite has been met, it will be necessary to build a standard way to introduce citations in a publication in such a way that aggregators and search engines can retrieve them and make good use of them: for instance by creating links between publications that would not have been possible otherwise.

In addition to the primary aspects connected to having high quality citations for data used in research and publications as credit, attribution, research verifiability etc., we should also consider the value of citations in increasing the discoverability of data and thus its reusability. Often a citation is the primary information source since the metadata for the resource is not always visible in any of the major metadata catalogues. In the Social Sciences well-structured catalogues have been developed for a long time as many datasets consist of a multitude of topics not covered by the dataset title.

Indeed, the title of the task in the SSHOC project is "Making data findable by being Citable". This clearly requires having proper descriptive metadata e.g. Creator, Organisation, Creation date, etc. as well as a PID for the data which can be used for data mining purposes (e.g. Crossref, OpenAIRE). Another way to increase discoverability is to enable the use of citations for semantic and/or web annotation. Groups of knowledgeable users are able to enrich the information available about the cited data by tagging the cited data resources with comments, terms from domain specific vocabularies, or even relations with other resources. The use of stable unique identifiers for the resource and its parts are essential for the sharing and discoverability of such annotations. The most popular model for this is standardised by W3C[54], but is not always sufficient[55] .

Concerning annotations, almost two thirds of libraries annotate data (e.g. by adding metadata or keywords) on a regular basis and almost half do the same with cleaning data. Editing of data (e.g., by adding mark-up to a document) is less popular. More than a third say they never do this (see graph on p. 16)." Europe's Digital Humanities landscape - A Report from LIBER's Digital Humanities & Digital Cultural Heritage Working Group[56]".

Another totally different approach to using potential actionability for data is proposed by the FAIRMetrics[57] Group: in this vision actionability means "measuring". This work, based on the GO-FAIR initiative, aims to measure the "FAIRNess" of data automatically by testing different criteria on line to evaluate a maturity indicator[58]. These results could be used to carry out some assessment at the data set level and also to evaluate repositories themselves.

[54] The Web Annotation Data Model (W3C 2017a, 2017b)

[55] Facilitating Fine-grained Open Annotations of Scholarly Sources. / Boot, P.; Haentjens Dekker, R.; Koolen, Marijn; Melgar, Liliana. https://pure.knaw.nl/portal/en/publications/facilitating-finegrained-open-annotations-of-scholarly-sources(c4fda9e2-bb42-4e8a-9c35-1361b89e0ee6).html
2017. Abstract from Digital Humanities 2017, Montreal, Canada.

[56] Europe's Digital Humanities landscape https://doi.org/10.5281/zenodo.3247285

[57] FAIRMetrics https://github.com/FAIRMetrics/

[58] FAIRMetrics/Metrics: FAIR Metrics, Evaluation results, and initial release of automated evaluator code https://zenodo.org/record/1305060#.XT_dzFDgrOQ

4. Making Data Actionable

In the previous section, we stressed the need to identify a dataset correctly. Of course, even if it is mandatory, it is not sufficient: we want to enrich and link datasets by using the content and to automate the process as much as possible.

"Actionability" is not easy to define, but in the specific context of data citation, it can be considered as the need to have usable information associated with data that is usable by both humans and machines. Another important part of actionability is to be able to retrieve this information in a consistent way, or even better in a standardized way. Therefore, it will be necessary to work on different topics to implement what we called "actionability": the information associated with datasets, a "protocol" to retrieve this information but also a way to link, and thereby enrich, this information.

For our purposes, the term 'actionable' is most relevant in the context of a reference embedded in a document or a bookmark to a resource on the internet. An end-user can activate such references for instance by mouse-clicking on the embedded reference, whereafter the referenced resource is visualised, or otherwise activated depending on its data-type. The simplest forms are URI's embedded in a web-page: much more complex embedded objects and activation procedures exist, but are usually application specific. Web browsers support a limited number of (configurable) action patterns for embedded objects e.g. mime-type client application mappings, the minimal pattern being opening new web-pages when a user clicks on an embedded URI. In the context of citations of publications, the obvious actionability is the download and/or visualisation of the paper in question, while in the case of data-set citations, the function is maybe somewhat less evident, but can minimally include visualisation of the data-set's metadata or its landing page.

Human actionability

The above describes actionability within software UI contexts i.e. human actionability.

Within SSH, an important example of the power of human actionability of data citations and data references, and that can be provided in a distributed fashion is the CLARIN Language Resource Switchboard[59]. When selecting a resource URI in a Switchboard implementing tool, the Switchboard offers users a list of services to choose from to process or visualise specific data-types. While the functionality the Switchboard offers is similar to what is offered by many computer operating systems when the user selects a file of a specific data-type, but the Switchboard offers this as a separate service available within all applications that choose to implement it. Generalizing the Switchboard is part of SSHOC project task 3.6 and the EOSC-hub project[60]

Machine actionability

There is however also 'machine actionability' when software agents are able to analyse data structures and identifiers and take specific actions when they encounter the right triggers. e.g. web-pages containing the right metadata attributes will trigger specific indexation options when a web-crawler passes by. In the context of

[59] The Language Resource Switchboard, Zinn, Claus, Computational Linguistics 2018 vol. 44 num. 4, p631-639, retrieved from https://doi.org/10.1162/coli_a_00329
[60] EOSC-Hub https://www.eosc-hub.eu

citations, the Event data project, detailed below, of cross-ref and Datacite is a good example of how citations become machine actionable with respect to mining publications for relations with data-sets.

Recently, Google has implemented a new tool, "Google Data Search"[61]. In order to make data discoverable, it is recommended to provide descriptive metadata in different kind of formats and structures[62] such as "schema.org" embedded in a web page or W3C DCAT vocabulary[63]. This approach can serve as a model to propose different flavors of structuring for SSH research data sets. A point to be noted is that this tool is linked with Google Scholar, sometimes in an inappropriate way as the link is based mostly on the name of the data set, but it is improving regularly. You can see an example by searching "Corpus de la Parole" in Google Data Search[64] and see which publications from Google Scholar are (supposedly) linked to this corpus[65].

Another way of using metadata is provided for instance by the different APIs[66] in Crossref. The results of the queries are returned in different formats, such as classical structured XML, UNIXSD or UNIXREF. These APIs can be used to achieve diverse goals, from text mining for documentalists, creation of references for researchers and accounting for funders.

A relatively new service is also proposed by DataCite and Crossref: Event Data[67]. The idea is to provide links between DOIs coming from Crossref, Datacite or other DOI registration agencies to give a fuller picture of the provenance, context etc. associated with the dataset. A list of possible relationships between different DOIs can be found on the DataCite web site[68]. The service is called "Event" in order to convey the dynamic side to theses links. The data itself come from different sources ranging from Datacite and Crossref to Twitter and Wikipedia. The main idea is that the data are no longer confined to repositories and then "invisible". On the contrary, data are now the subject of multiple references on the web at large such as citation in a blog, sharing on social networks, comments, links, bookmarks etc.: all these references are considered as "events" that are stored and to which this service gives access. It is important to note that all events, even if they are stored without being processed, are associated with information about their provenance, context, dates etc. The technology used to associate these events to a DOI is based on a classical triple from Semantic Web Technologies. However, the idea is not to provide a graph, but rather to enable users to build their own graphs by using the API. According to the service provider, it could be used for a great variety of services from metrics for funders and publishers to research purposes such as building graphs of related data.

61 Google Data Search https://toolbox.google.com/datasetsearch and https://en.wikipedia.org/wiki/Google_Dataset_Search
62 Google Dataset recommendations https://developers.google.com/search/docs/data-types/dataset
63 W3C DCAT Vocabulary https://www.w3.org/TR/vocab-dcat/
64 Google Data Toolbox https://toolbox.google.com/datasetsearch/search?query=cocoon%20huma-num&docid=2NAn%2FfIEOieSTPbcAAAAAA%3D%3D
65 Google Scholar Links https://scholar.google.com/scholar?q=%22corpus%20de%20la%20parole%22
66 Crossref REST API https://www.crossref.org/services/metadata-delivery/rest-api/
67 Event Data https://www.eventdata.crossref.org/
68 Event Data relationships https://support.datacite.org/docs/eventdata-guide#section-relation-type-id

Lastly, we can also mention the new features introduced in Zenodo, which is still in an early phase of development, to count citations[69]. There will be a dedicated part for Dataset, and it will be based partly on Datacite and Crossref work already mentioned in this document.

Organisations involved with storing, managing and publishing research data mostly use data repository systems as DSpace, Fedora, DataVerse etc. Although these systems also permit direct access to the data objects, the organisations managing the repositories prefer to promote access via a so-called landing page which lists descriptive and technical metadata for the resource, licensing and usage conditions, and a link to the data resource itself (bitstream). Offering direct access to resources is not popular and often even impossible for legal and branding reasons and requires a user to make decisions on the basis of the landing-page content. This pattern of providing access through landing pages is often also followed in cases where there is no absolute necessity for human mediation exists.

In such systems that have been set up primarily for the use of humans using a web browser, who follow a PID to a landing-page, locate the URI to the actual data on that page, and then access the data, having a software agent perform that task in a similar fashion is quite complicated. However, some strategies can help. First there is the option of 'standardizing the data landing page' facilitating easy location of the data link. Within a specific repository type, standardization should prove relatively easy. Another option is to make use of "http signposting"[70] which makes it possible to automatically resolve resources behind landing pages.

This still leaves certain complex matters unaddressed before processing the data, such as identifying the type of the data, determining the available data formats, evaluating the data license and the possible use constraints, authentication and authorization for protected resources. However, solutions are available and standardization initiatives are underway for instance in the Social Sciences. And most importantly, many resources are in principle open available.

From a technical point of view, implementing machine-actionable identifiers requires a complex software infrastructure that includes: communication protocols, data and metadata parsers, efficient authentication/authorization management systems, query management systems, etc.

The RDA Working Group on Data Citation[71] produced a set of valuable recommendations[72], covering versioning, PID issuing and the use of landing-pages to bundle relevant information. Currently, within RDA, GEDE is working on ways to build on the Digital Object (DO) model and enhance the functionality of PIDs, including their machine actionability. GEDE has formed the GEDE Citation Topic Group[73] to produce a consensual synthesis of existing recommendations, focussing both on citing from publications and data-to-data citing.

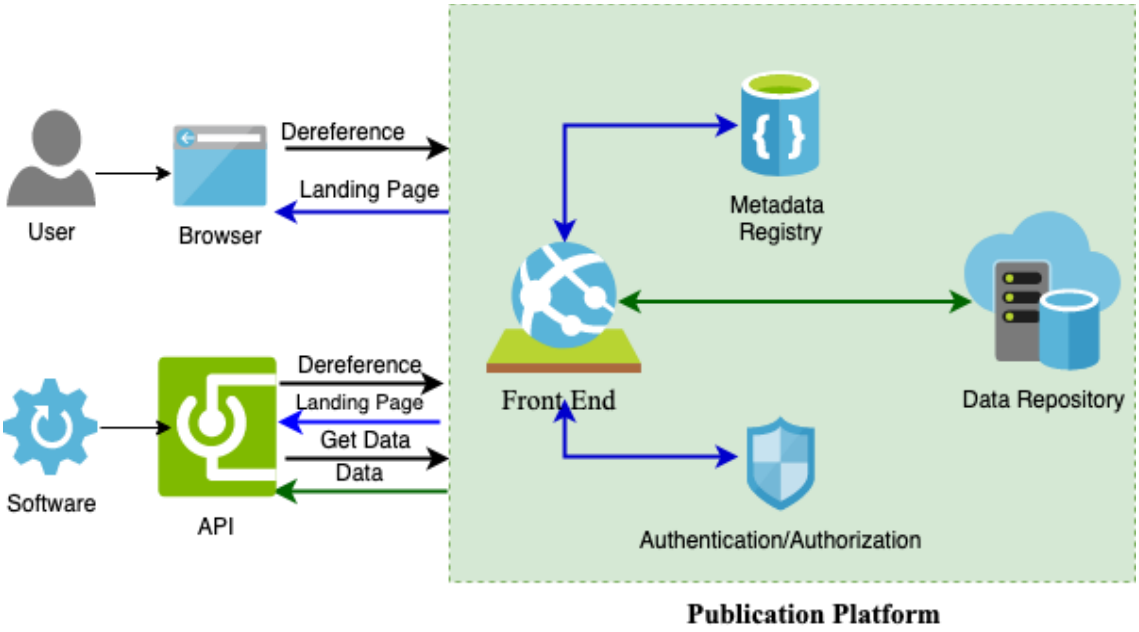
69 Data citation in Zonodo https://help.zenodo.org/

70 Klein, M., Shankar, H., and Van de Sompel, H. (2018) Signposting for Repositories. Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018. https://doi.org/10.1145/3197026.3203879

71 RDA Data Citation recommendation https://rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html

72 Andreas Rauber; Ari Asmi; Dieter van Uytvanck; Stefan Proell (2015): Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). DOI: 10.15497/RDA00016

73 RDA GEDE https://rd-alliance.org/group/gede-group-european-data-experts-rda/wiki/gede-citation-topic-group



Identifier actionability

The image above shows the main components involved in implementing actionability for identifiers. In human-actionability a user interacts with the publication platform using a software agent (e.g. a browser) and obtains a landing page containing: metadata description of the dataset, security and license information, and a link to the actual dataset. In machine-actionability, a software module may use API to interact with the publication platform and obtain the dataset. In this case the link to the dataset is automatically individuated by parsing the landing page, this means that the agent should know how to parse the landing page, for instance that the page is formatted using Signposting patterns.

5. Motivating for data citation in the SSH

Many of the SSHOC partners have played important roles with respect to citation recommendations or standards. This ranges from participation in ISO-TC37 (CLARIN, DARIAH), to participation in relevant RDA working groups on PIDs and citation. Some partner infrastructures have dedicated committees and task forces on these subjects e.g. the CLARIN PID task force.

With respect to the role of SSHOC in this task we hope that this document can be a basis for a living document to guide SSH citation practices that will be expanded during the project. In addition to that, it will be necessary to take into account developments of this area that will emerge during the project duration: the topic is new to some but also quite dynamic and we can expect some moves, for instance by private publishers.

Data Management Motivators

A large group of organisations and initiatives can work as motivators for data management and data citation.

The European Open Science Cloud (EOSC)[74] is at present the most important initiative in Europe that supports data citation, whose strategic implementation plan[75] was published in July 2019. The EOSC has its portal[76] as the main entry point and "will offer 1.7 million European researchers and 70 million professionals in science, technology, the humanities and social sciences a virtual environment with open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines by federating existing scientific data infrastructures, currently dispersed across disciplines and the EU Member States."[77] It is a collaborative initiative that was proposed by the European Commission in 2016 and counts a number of doers and enablers, including the EOSC Declaration[78] signatories, the EOSC Board[79], the EOSC Executive Board[80], the EOSC working groups[81], the EOSC Stakeholders Forum[82], as well as several implementation projects, including the EOSC Secretariat[83] and the EOSC cluster projects (ESFRI thematic projects): SSHOC[84], ENVRI-FAIR[85], PANOSC[86], ESCAPE[87] and EOSC-Life[88], supported by the EOSC-hub[89]. The EOSC Declaration stipulated as a prerequisite for the realisation of the EOSC and its technical implementation that "a data citation system should be put in place to reward the provision of excellent open data. This will assist both the assessment of researchers and their projects, and help implementing the findability, accessibility, interoperability and reusability of research data". The FREYA[90] project is currently working on the development of a PID infrastructure intended for the EOSC.

National initiatives also work as motivators for data management and data citation, thanks to their close relationship with national funding agencies that can make demands with respect to data sharing and research reproducibility. Adopting National Plans on Data Management are of particular importance as they help shape and form European Initiatives in an effort to achieve standardisation. By developing concrete national plans that cater for the effective management of data and via the sharing of these plans and policies, these countries can serve as an inspiration, not only on a pan-European level, but for inspiring and motivating neighbouring countries, that have not yet adopted data management plans or that are lagging behind on developments. Good examples of this would are the Aporta Initiative[91] in Spain dedicated to promoting the opening of public

74 EOSC https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
75 EOSC Strategic Implementation Plan https://ec.europa.eu/info/publications/european-open-science-cloud-eosc-strategic-implementation-plan_en
76 EOSC Portal https://www.eosc-portal.eu/
77 EOSC Description https://www.eosc-portal.eu/about/eosc
78 EOSC Declaration https://ec.europa.eu/research/openscience/pdf/eosc_declaration.pdf#view=fit&pagemode=none
79 EOSC Board https://www.eosc-portal.eu/governance/eosc-board
80 EOSC Executive Board https://www.eosc-portal.eu/governance/executive-board
81 EOSC Working Groups https://www.eoscsecretariat.eu/eosc-working-groups
82 EOSC Stakeholders Forum https://www.eosc-portal.eu/governance/stakeholders-forum
83 EOSC Secretariat https://www.eoscsecretariat.eu/node
84 SSHOC Project https://www.sshopencloud.eu/
85 ENVRI-FAIR Project http://envri.eu/envri-fair/
86 PANOSC Project https://www.panosc.eu/
87 ESCAPE Project https://projectescape.eu/
88 EOSC-LIFE Project https://cordis.europa.eu/project/rcn/219199/factsheet/en
89 EOSC Hub https://www.eosc-hub.eu
90 Project FREYA https://www.project-freya.eu
91 APORTA Initiative https://datos.gob.es/en/about-aporta-initiative

information and development of advanced "data-base" services, the Public Data Policy[92] in France and the German National Data Portal[93] in Germany.

European and global data initiatives also play a significant role in motivating researchers and research data users in general. The Research Data Alliance (RDA) "builds the social and technical bridges to enable the open sharing and re-use of data",[94] while it provides a neutral venue for several stakeholder categories to come together using various channels, such as working groups, interest groups, recommendations and guidelines for its members to develop and adopt infrastructures for research data management. EUDAT "is a Service-oriented, Community driven, Sustainable and Integrated initiative".[95] It works towards its vision through a Collaborative Data Infrastructure (CDI)[96] and its network across 15 European nations. Another example of a motivator at the European level is OpenAIRE[97] with a wide spectrum of support, training and advocacy activities, especially through its National Open Access Desks (NOADs)[98]. Furthermore, OpenAIRE announced in 2018 its support[99] for I4OC[100], an initiative for Open Citations to promote the unrestricted availability of scholarly citation data.

Research and academic libraries serve as a very proximal driving force in supporting researchers to consider best practices for data citation. The traditional role of the librarian is rapidly shifting along with the scientific research developments. In a data-driven scientific research environment, academic librarians are taking on new roles and developing new skills in order to guide researchers through the sometimes labyrinthine scholarly publication process, data repositories and policies. This need has created new roles for librarians and now it is common practice for most university libraries to have specialized librarians (digital scholarship librarians, data librarians) dedicated to supporting researchers publish their work while remaining compatible with data citation and data curation principles.

Already in 2012, LIBER's working group on E-Science / Research Data Management published "10 recommendations for libraries to get started with research data management"[101], showing that the role of libraries in data citation and overall in data management includes training the researchers, research management support when it comes to grant submission, development of metadata and data standards, participation in institutional research data policy development, including resource plans and promoting research data citation by applying persistent identifiers to research data amongst others. In B. Schmidt et al.[102]

92 French Public Data Policy https://www.gouvernement.fr/en/public-data-policy
93 German National Data Portal https://www.govdata.de/
94 RDA Alliance https://www.rd-alliance.org/about-rda
95 EUDAT Foundation https://eudat.eu/what-eudat
96 EUDAT Collaborative Data Infrastructure https://www.eudat.eu/eudat-cdi
97 OpenAIRE https://www.openaire.eu/
98 OpenAIRE National Open Access Desks https://www.openaire.eu/contact-noads
99 OpenAIRE support t to I4OC https://www.openaire.eu/openaire-is-proud-to-support-the-new-initiative-for-open-citations-i4oc
100 Initiative for Open Citation https://i4oc.org/
101 LIBER Research Data Group https://www.openaire.at/fileadmin/user_upload/openaire/LIBER_The_research_data_group_2012_v7_final.pdf
102 Librarians' Competencies Profile for Research Data Management, 2016, https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf

the core competencies of librarians to be able to support research data management, including at least a basic understanding of the disciplinary landscape, norms, and standards, are also complemented by advocacy and support for managing data. When it comes to data citation, a thorough knowledge of data citation and referencing practices, as well as best practices for data structures, types, formats, vocabularies, ontologies and metadata is a requirement.

The University libraries as a natural hub for researchers and scientists take on the role of providing them with the necessary tools and skills for the efficient management and dissemination of their data in order to ensure the transparency and interoperability of the research process as well as making their data F.A.I.R. [Findable, Accessible, Interoperable, Re-usable]. Libraries now offer templates for successful management plans along with workshops and continuous training for the researchers. This service provided by research and academic libraries is of particular importance as it provides the researchers with invaluable insight during all phases of the research cycle explaining in detail the data management principles.

LIBER's Research Data Management Working Group[103] has also worked extensively on creating a DMP Catalogue which is available on the LIBER Website[104] and provides insight and inspiration for researchers across a wide variety of disciplines. The Working Group has also been very active in hosting webinars and workshops in Finding and Reusing Research Data, in researching best practices for Data Repositories in academic institutions and training the libraries to effectively support the data science needs of their researchers. As early as 2011, the Working Group published "Ten recommendations for libraries to get started with research data management"[105] which remains highly relevant today as it highlights among others the need to promote research data citation by applying persistent identifiers to research data and the support of the life cycle for research data by providing services for storage, discovery and permanent access.

Effective and successful data citation that would result in better use of research data, following FAIR principles, starts at the beginning of the scientific research process, which is why the libraries and librarians need to train and motivate researchers throughout this process.

By providing training for researchers on DMP, academic libraries ensure that best practices for data citation are documented and serve as a guide during the research cycle.

Research funders are a pivotal stakeholder in achieving Open Science. Through their mandates, they can in this sense make data management plans and data citation a standard procedure in the data cycle, by making it a requirement for the grant. In this sense, it is common to see research funders providing guidelines for a sound

103 LIBER Research Data Management Working Group https://libereurope.eu/strategy/research-infrastructures/rdm/
104 LIBER DMP https://libereurope.eu/dmpcatalogue/
105 LIBER "Ten recommendations for libraries to get started with research data management" https://libereurope.eu/wp-content/uploads/The%20research%20data%20group%202012%20v7%20final.pdf

data management plan at the time of the proposal (European Commission,[106] Wellcome Trust,[107] Gates Foundation,[108] etc.)

6. Conclusion and future work

The SSHOC project represents a unique opportunity to make a number of observations and recommendations for a common data citation approach covering a broad selection of the SSH communities. As has been established in this document much work with respect to data citations has already been done in many different research disciplines including the SSH. However in recent years the unique aspects of data citations versus the classical citation of publications and also the nature of typical SSH data-types have resulted in new opportunities and infrastructure developments for supporting and improving the reuse and discoverability of data, which have been described in this deliverable and should be supported by our recommendations.

It was essential to take into account new types of data that have emerged in recent years, and are widely used now in the SSH landscape, such as dynamic data coming for instance from social networks like Twitter. Likewise, another interesting source for the SSH communities which is the flow of comments, and annotations from people on the web or done by machines etc. that are generated by "users", regardless of the definition adopted for a user.

This leads to the need to provide for implementations that can manage consistent dynamic data citation and different types of annotations including leveraging the wisdom of expert and non-expert users through "semantic annotation".

A point that was already referred to and will become crucial in the near future is the ability to link datasets and more traditional publications such as papers. Some benefits should be expected from this process for SSH: better reusability of datasets, better reproducibility of research and clearer visibility for funding agencies (e.g. PID graph developed by the FREYA project). Several different initiatives already exist, but a role for this task should be to identify what is really usable and propose a realistic plan of action to build the "data-publication" link.

As mentioned, there are already many existing initiatives and technologies for data citation and many initiatives are still being developed for specific disciplines: for instance, the "Linguistic Data Interest Group[109]" from RDA. Even if they are not always well suited to SSH needs, we can't ignore them and reinvent the wheel by starting from scratch. Therefore, our approach for the recommendations given in this document are practical rather than highly technical. We see a diverse audience for these recommendations such as infrastructure engineers, data managers and of course researchers and research project teams in general.

[106] EC Guidelines for Data Management https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
[107] Wellcome Trust Guidelines for Data Management https://wellcome.ac.uk/funding/guidance/how-complete-outputs-management-plan
[108] Gates Foundation Guidelines for Data Management https://gatesopenresearch.org/documents/2-46
[109] RDA Linguistic Data Interest Group https://www.rd-alliance.org/groups/linguistics-data-ig

With respect to these goals, a good basis for deciding what should be essential for also SSH data citations is to be, as far as possible, compliant with the "Data Citation Principles" defined by Force11 which are summarised as[110] "Importance", "Credit and Attribution", "Evidence", "Unique Identification", "Access", "Persistence", "Specificity and Verifiability" and "Interoperability and Flexibility".

When considering the investigated citation practices and technologies from a practical point of view there are a number of aspects that we want to underline as beneficial for the SSH:

1. the use of explicit versioning policies requiring an explicit specification of the data object version or the lack thereof.
2. the use of a "tombstone" landing page associated with PIDs for deleted data objects
3. support for citing parts of complex objects i.e. using PIDs that support part-identifiers or some other suitable technology, that is able to efficiently access parts of larger objects.
4. Some citation metadata seem to be overly extensive, though we cannot judge the absolute necessity of such a practice. However, we do want to recommend using DataCite citation metadata as a minimal set and keeping the citation metadata manageable and with a predictable structure.
5. increase the actionability of citations and landing pages: using actionable identifiers and well - structured and informative landing pages. This is also consistent with the DataCite landing page recommendations[111]
6. It would be useful to implement interoperability services in the publication platform: for instance, providing standard harvesting protocols that could improve the possibility of sharing metadata with other platforms, providing a well-documented API to access data can simplify machine-actionability for identifiers, etc

All these recommendations can be summarized in "achieving greater visibility" which will raise awareness of what SSH are doing and what type of data are being used.

Future Work

In this document, we have listed different approaches to making data actionable. This task, for the future, will be defined in cooperation with other SSHOC tasks to decide which approaches would be suitable for the SSH. We also see a need to continue monitoring initiatives concerning data citation at large since many concurrent initiatives are currently developing around research data management, from Data Management Plans to citation, usage measurement etc. Because of this dynamic data management landscape, this deliverable should be seen as a basis for a living document for guiding SSH citation practices that can be expanded during the project.

Although we have not listed all the existing technologies and recommendations, we note that an important feature is missing in almost every case, namely the ability to "interact" with the creator and other users of the data set. An additional recommendation would be to add "social possibilities" to repositories and citations to provide classical services such as comments, notifications etc. In order to do that, it will be necessary to define a specific set of metadata, associated with vocabularies and ontologies and find simple protocols to exchange

[110] Joint Declaration of Data Citation Principles https://www.force11.org/datacitationprinciples
[111] https://support.datacite.org/docs/landing-pages

such information. With regard to the use of identifiers in particular, it seems that a good approach would be to combine several technologies. The idea would be, for instance, to use "handle technology" to provide a direct access to data and rather " DOI technology" to access metadata: there already exist proposals associated with DOIs, such as "DOI Citation Formatter[112]", that it would be useful to be able to reuse.

In the context of the SSHOC project itself, task 3.4 plans to liaise and collaborate with all other SSHOC initiatives where the ideas and recommendations set out in this deliverable are relevant. But in particular we see relevance with:

- SSHOC task 3.6, that aims to generalise the CLARIN Language Resource Switchboard into a SSH infrastructure component, and that should be an excellent way to make citations Human actionable
- SSHOC task 5.2, that will develop a SSHOC DataVerse repository version. DataVerse already supports creating citations, but we will investigate the possibility of introducing semantic annotation options
- SSHOC WP6 is planning training events on a number of data management issues including citation.
- SSHOC is relying on the SSHOC stakeholder infrastructures to contribute key infrastructure components to a common SSH infrastructure. Where practical, we will discuss where practical, the implementation and integration of our citation recommendations and citation infrastructure ideas

More generally, we also need to identify which communities are actually using data citation and see how the SSH compare and could accommodate. Firstly, we plan to select conferences, like RDA, to present our case and discuss our ideas. Secondly, we should also think about organizing workshops on behalf of SSHOC to discuss with different communities and stakeholders and add our findings to this document.

An interesting concept that appeared during the writing of this document is the notion of "Citation Infrastructure": this could be a repository associated with annotation features, discoverability etc. Even if this concept needs to be clarified it is a good base for the task's future work if we can connect it to devising a prototype, relying as much as possible on existing implementations, which would implement our recommendations and suggested infrastructure. For instance, this prototype could implement a recommendation for a landing page containing well-structured information inside (e.g. RDFa) and how to "dereference" it based on the work already done by DataCite as a minimum. The SSHOC task 3.4 partners already control some infrastructure components that will be helpful in this matter.

In this regard, SSHOC also represents a unique opportunity since many of the key players are involved, which guarantees that exchanges and experiments will be efficient.

[112] DOI Citation Formatter https://citation.crosscite.org/

7. References

Articles and Reports

- Developing Data Attribution and Citation Practices and Standards: Summary of an International Workshop - https://www.nap.edu/read/13564/chapter/1
- State of the art report on open access publishing of research data in the humanities from Has (Humanities At Scale) project - https://hal.archives-ouvertes.fr/halshs-01357208
- CODATA report - http://datascience.codata.org/articles/abstract/10.2481/dsj.OSOM13-043
- http://www.nap.edu/catalog/13564/for-attribution-developing-data-attribution-and-citation-practices-and-standards
- The Problem of Citation in the Digital Humanities / Jonathan Blaney https://www.dhi.ac.uk/openbook/chapter/dhc2012-blaney
- Data citation practices across earth, life, social sciences and humanities - opportunities for OpenAIRE https://zenodo.org/record/54570/files/Data-Citation.pdf
- Chen, Xiaoli. (2017, October). Where are we with Data Citation. Zenodo. http://doi.org/10.5281/zenodo.1004744
- Graef, Florian, & McEntyre, Jo. (2016). OpenAIRE2020 D7.5 – Data citation standards and index requirements. Zenodo. https://doi.org/10.5281/zenodo.1257267
- Dorch, S. B. F. (2012, July 5). On the Citation Advantage of linking to data. Zenodo. http://doi.org/10.5281/zenodo.6772
- Parsons, Mark A., & Fox, Peter A. (2014). Why Data Citation Currently Misses the Point. Zenodo. http://doi.org/10.5281/zenodo.1241521
- Hourclé, J., Chang, W., Linares, F., Palanisamy, G., & Wilson, B. (2012). Linking Articles to Data. Zenodo. http://doi.org/10.5281/zenodo.13802
- Kristian Garza. (2018, October). Exploring data citation in Crossref and DataCite's Event Data service. Zenodo. http://doi.org/10.5281/zenodo.1472279
- Parsons, Mark A., Duerr, Ruth E., & Jones, Matthew B. (2019). The History and Future of Data Citation. Zenodo. http://doi.org/10.5281/zenodo.2619468
- Hourclé, J. (2014). Data Citation in Astronomy. Zenodo. http://doi.org/10.5281/zenodo.10505
- Wright, C., Hodson, S., Deventer, M. van ., Selematsela, D., Lötter, L., Vahed, A., … Walt, I. van . der . (2016, January). CODATA Data Citation Workshop, South Africa: Presentations. Zenodo. http://doi.org/10.5281/zenodo.44946
- Rueda, Laura. (2016, November). Making research better by enabling people to find, share, use and cite data. Zenodo. http://doi.org/10.5281/zenodo.168213
- Hourclé, Joseph. (2015, April). Data Realities: Citation Equals Funding. Zenodo. http://doi.org/10.5281/zenodo.2654574
- LIBER Research Data Group https://www.openaire.at/fileadmin/user_upload/openaire/LIBER_The_research_data_group_2012_v7_final.pdf
- Librarians' Competencies Profile for Research Data Management, 2016, https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf

- The Anatomy of a Data Citation: Discovery, Reuse, and Credit https://academiccommons.columbia.edu/doi/10.7916/D8MW2STM
- Credit data generators for data reuse https://www.nature.com/articles/d41586-019-01715-4
- Data Citation as a computational problem https://frew.eri.ucsb.edu/private/preprints/bdf-cacm-data-citation.pdf
- Mary Vardigan: Data Citation for the Social Sciences, https://www.nap.edu/read/13564/chapter/9
- Sieber, Joan E. & Bruce E. Trumbo (1995). "(Not) giving credit where credit is due: Citation of data sets". Science and Engineering Ethics 1: 1, 11–20.
- Dodd, Sue A. (1979). "Bibliographic References for Numeric Social Science Data Files: Suggested Guidelines". Journal of the American Society for Information Science 30: 2, 77–82.
- Marwick, B. and S. E. Pilaar Birch (2018) A Standard for the Scholarly Citation of Archaeological Data. *Advances in Archaeological Practice* https://doi.org/10.1017/aap.2018.3
- Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., … Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data, 3, 160018. https://doi.org/10.1038/sdata.2016.18
- «Presentation of Data Papers», Cybergeo : European Journal of Geography [En ligne], Data papers, http://journals.openedition.org/cybergeo/28922
- Brill Research Data Journal https://brill.com/view/journals/rdj/rdj-overview.xml?lang=en
- The Web Annotation Data Model (W3C 2017a, 2017b)
- Facilitating Fine-grained Open Annotations of Scholarly Sources. / Boot, P.; Haentjens Dekker, R.; Koolen, Marijn; Melgar, Liliana. https://pure.knaw.nl/portal/en/publications/facilitating-finegrained-open-annotations-of-scholarly-sources(c4fda9e2-bb42-4e8a-9c35-1361b89e0ee6).html
- Europe's Digital Humanities landscape https://doi.org/10.5281/zenodo.3247285
- The Language Resource Switchboard, Zinn, Claus, Computational Linguistics 2018 vol. 44 num. 4, p631-639, retrieved from https://doi.org/10.1162/coli_a_00329
- Klein, M., Shankar, H., and Van de Sompel, H. (2018) Signposting for Repositories. Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries, JCDL 2018. https://doi.org/10.1145/3197026.3203879
- Andreas Rauber; Ari Asmi; Dieter van Uytvanck; Stefan Proell (2015): Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). DOI: 10.15497/RDA00016 https://zenodo.org/record/1406002#.XG2FVOhKhaQ

Recommendations and Guidelines

- Force 11 https://www.force11.org/datacitationprinciples
- DataVerse https://dataverse.org/best-practices/data-citation
- Data On the Web Best Practices https://www.w3.org/TR/dwbp/
- SHCOLIX http://www.scholix.org/; https://www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg
- Librarians' Competencies Profile for Research Data Management, 2016, https://www.coar-repositories.org/files/Competencies-for-RDM_June-2016.pdf
- The FAIR Guiding Principles for scientific data management and stewardship https://www.nature.com/articles/sdata201618

- Task Group on Data Citation Standards and Practise of Cite, Out of Mind: The Current Sices, C.-I., 2013. Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal, 12, pp.CIDCR1–CIDCR7. DOI: http://doi.org/10.2481/dsj.OSOM13-043
- APA, MLA style guides MLA Formatting Guide https://owl.purdue.edu/owl/research_and_citation/mla_style/mla_style_introduction.html
- ISO24615 PISA Persistent Identification and Sustainable Access of Language Resources (2011)
- AILLA citation guidelines https://ailla.utexas.org/site/rights/citation
- ANDS https://www.ands.org.au/working-with-data/citation-and-identifiers/data-citation
- DataCite https://datacite.org/cite-your-data.html Metedata Working Group https://schema.datacite.org/ http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf
- RDA Data Citation Working group / recommendations, https://rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html https://www.rd-alliance.org/group/data-citation-wg/post/small-chapter-dynamic-data-citation-published
- ICPSR Data Citation Recommendations https://www.icpsr.umich.edu/icpsrweb/content/datamanagement/citations.html
- LIBER Research Data Management Working Group https://libereurope.eu/strategy/research-infrastructures/rdm/
- LIBER DMP https://libereurope.eu/dmpcatalogue/
- EC Guidelines for Data Management https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
- Wellcome Trust Guidelines for Data Management https://wellcome.ac.uk/funding/guidance/how-complete-outputs-management-plan
- Gates Foundation Guidelines for Data Management https://gatesopenresearch.org/documents/2-46
- Zotero bibliography https://www.zotero.org/groups/514443/open_data_citation_for_social_science_and_humanities? Data Citation in Zenodo "Citations" https://help.zenodo.org/ - Use of https://asclepias-broker.readthedocs.io ?
- COST ENRESSH Action https://enressh.eu/wp-content/uploads/2018/04/WG3_Ljubljana_Istenic_etal.pdf
- https://enressh.eu/links_and_literature/presentations/
- Using The Wayback Machine https://help.archive.org/hc/en-us/articles/360004651732-Using-The-Wayback-Machine
- TEI Guidelines https://tei-c.org/guidelines/
- DASISH D 5.1 Trust and PID Services Federation Report https://dasish.eu/publications/projectreports/DASISH-D5_1_BS_version141106.pdf
- RDA Data Citation of Evolving Data https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf
- CTS URN http://www.homermultitext.org/hmt-docs/cite/cts-urn-overview.html
- CESSDA Citing your data https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/6.-Archive-Publish/Publishing-with-CESSDA-archives/Citing-your-dat
- CESSDA Access Use and Cite your data https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/Access-use-and-cite-data

- CESSDA ERIC Persistent Identifier Policy http://multiweb.gesis.org/csaw/#!Detail/cessda-eric/0047
- Europeana Resource Citation task force https://pro.europeana.eu/project/resource-citation-object-identity-standardization
- European Persistent Identifier Consortium https://www.pidconsortium.eu
- IASSIST Data Citation Resources https://iassistdata.org/community/data-citation-ig/data-citation-resources
- SHARE Citation Requirements http://www.share-project.org/data-access/citation-requirements.html
- RDA GEDE Webworkshop https://www.rd-alliance.org/groups/gede-group-european-data-experts-rda
- RDA Persistent identifiers: Consolidated assertions https://www.rd-alliance.org/group/data-fabric-ig/outcomes/persistent-identifiers-consolidated-assertions
- CLARIN PID policy from https://www.clarin.eu/node/3965
- DataCite Metadata Schema https://schema.datacite.org/ and http://schema.datacite.org/meta/kernel-3/doc/DataCite-MetadataKernel_v3.1.pdf
- DataCite Repository Software Integrations https://support.datacite.org/docs/repository-software-integrations
- Data Citation in DataVerse https://dataverse.org/best-practices/data-citation
- Google Dataset recommendations https://developers.google.com/search/docs/data-types/dataset
- Event Data relationships https://support.datacite.org/docs/eventdata-guide#section-relation-type-id
- Data citation in Zonodo https://help.zenodo.org/
- RDA Data Citation recommendation https://rd-alliance.org/group/data-citation-wg/outcomes/data-citation-recommendation.html
- RDA Linguistics Data Interest Group https://www.rd-alliance.org/groups/linguistics-data-ig

Projects, Institutions and Initiatives

- UK Data archive http://data-archive.ac.uk/conditions/mciting-data
- IANUS project https://www.ianus-fdz.de/attachments/download/659/Praesentation_2014-09-13_Final_kurz.pdf
- DANS https://dans.knaw.nl/en
- DDI Alliance http://www.ddialliance.org/
- OLAC http://www.language-archives.org/
- TEI Consortium https://tei-c.org/
- THOR https://project-thor.eu
- ODIN https://odin-project.eu/project-outputs/deliverables/
- DASISH https://dasish.eu/deliverables/
- FREYA https://www.project-freya.eu/
- ROR Community https://ror.community/about/
- ISSN https://road.issn.org/
- Initiative for Open Citation https://i4oc.org
- JATS Data Citation https://jats4r.org/data-citations
- Scientific Data (Nature) https://www.nature.com/sdata/
- GO-FAIR Initiative https://www.go-fair.org/
- RDA https://rd-alliance.org

- DASISH https://dasish.eu
- Open Research Data Pilot https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
- UK Web Archive EU referendum https://www.webarchive.org.uk/cy/ukwa/collection/649
- Google Data Search https://toolbox.google.com/datasetsearch and https://en.wikipedia.org/wiki/Google_Dataset_Search
- FAIRMetrics https://github.com/FAIRMetrics/ and https://zenodo.org/record/1305060#.XT_dzFDgrOQ
- Mendely Data https://data.mendeley.com
- EOSC-Hub https://www.eosc-hub.eu
- Event Data https://www.eventdata.crossref.org/
- RDA GEDE https://rd-alliance.org/group/gede-group-european-data-experts-rda/wiki/gede-citation-topic-group
- EOSC https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud
- EOSC Strategic Implementation Plan https://ec.europa.eu/info/publications/european-open-science-cloud-eosc-strategic-implementation-plan_en
- EOSC Portal https://www.eosc-portal.eu/
- EOSC Description https://www.eosc-portal.eu/about/eosc
- EOSC Board https://www.eosc-portal.eu/governance/eosc-board
- EOSC Executive Board https://www.eosc-portal.eu/governance/executive-board
- EOSC Working Groups https://www.eoscsecretariat.eu/eosc-working-groups
- EOSC Stakeholders Forum https://www.eosc-portal.eu/governance/stakeholders-forum
- EOSC Secretariat https://www.eoscsecretariat.eu/node
- SSHOC Project https://www.sshopencloud.eu/
- ENVRI-FAIR http://envri.eu/envri-fair/
- PANOSC Project https://www.panosc.eu/
- ESCAPE Project https://projectescape.eu/
- EOSC-LIFE Project https://cordis.europa.eu/project/rcn/219199/factsheet/en
- APORTA Initiative https://datos.gob.es/en/about-aporta-initiative
- French Public Data Policy https://www.gouvernement.fr/en/public-data-policy
- German National Data Portal https://www.govdata.de/
- RDA Alliance https://www.rd-alliance.org/about-rda
- EUDAT Foundation https://eudat.eu/what-eudat
- EUDAT Collaborative Data Infrastructure https://www.eudat.eu/eudat-cdi
- OpenAIRE https://www.openaire.eu/
- OpenAIRE National Open Access Desks https://www.openaire.eu/contact-noads

Others

- OpenAire Data Reporsitories https://www.openaire.eu/find-trustworthy-data-repository-open-data-pilot-requirements
- DOI Citation Formatter https://citation.crosscite.org/
- Winter School HaS https://datacite.hypotheses.org/

- Data Citation Index (Web of Science) https://clarivate.com/products/web-of-science/web-science-form/data-citation-index/
- InFoLIS http://infolis.github.io/
- Elixir citation with schema.org https://docs.google.com/document/d/1XuQ0GKkIcYD9Vyc1zIeVDyC_abfti1yGqcDYMyR71iQ/edit
- Crossref REST API https://www.crossref.org/services/metadata-delivery/rest-api/
- Data Verse Dataset management http://guides.dataverse.org/en/latest/user/dataset-management.html
- Zenodo Meta Data Extraction https://gist.github.com/xiaom/106b9d111726cc99017b7efe7aead01c
- OpenAire ScholeXplorer http://scholexplorer.openaire.eu/index.html#/
- Cool URIs for the Semantic Web http://www.w3.org/TR/2008/NOTE-cooluris-20081203/
- EPIC API Part Identifier https://doc.pidconsortium.eu/guides/api-partial/
- Mendeley API https://dev.mendeley.com/
- W3C DCAT Vocabulary https://www.w3.org/TR/vocab-dcat
- Google Data Toolbox https://toolbox.google.com/datasetsearch/search?query=cocoon%20huma-num&docid=2NAn%2FfIEOieSTPbcAAAAAA%3D%3D
- Google Scholar Links https://scholar.google.com/scholar?q=%22corpus%20de%20la%20parole%22
- Crossref REST API https://www.crossref.org/services/metadata-delivery/rest-api/
- OpenAIRE support to I4OC https://www.openaire.eu/openaire-is-proud-to-support-the-new-initiative-for-open-citations-i4oc