SSHOC

social sciences & humanities open cloud

Recommendations for FAIR Data Citation in the Social Sciences and Humanities

September 2021

SSHOC, "Social Sciences and Humanities Open Cloud", has received funding from the European Union's Horizon 2020 project call H2020-INFRAEOSC-04-2018, Grant Agreement #823782.

Recommendations for FAIR Data Citation in the Social Sciences and Humanities

Introduction

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

This document is intended for the various actors involved in the creation, treatment, curation, and use of research data in the SSH. These recommendations were conceived according to the FAIR principles, that is Findable, Accessible, Interoperable, and Reusable. This document proposes a set of simple recommendations on how to cite data in order to foster the visibility of research data, give credit to their creators and enhance trust in the research process.

The following recommendations are adapted for data citation within the SSH communities from the "Data Citation Principles" defined by Force11, the community that developed the FAIR Principles.: "Importance", "Credit and Attribution", "Evidence", "Unique Identification", "Access", "Persistence", "Specificity and Verifiability" and "Interoperability and Flexibility". They complement the SSHOC Inventory of SSH citation practices and draw inspiration from the session on FAIR Data-Citation for Social Sciences and Humanities, which took place during the Realising the EOSC event in November 2020.

An aspect identified in the aforementioned session, which encompasses the following recommendations, is the lack of incentives for all stakeholder categories involved in the research data lifecycle, especially for researchers. It is important to set up a rewards system for data creators (e.g., using the contributor taxonomy proposed by NISO). Such an action aims to not only provide motivation, but also to reward and promote the value of best practices. A second aspect to be further explored is the importance of data publication (e.g.,data papers and Data Journals). A third aspect would be to take inspiration from those fields at the forefront of standardized data publication, such as Astronomy, with an existing culture of data sharing.

Nicolas Larrousse, CNRS & Edward J. Gray, CNRS

Use cases for FAIR Data Citation in SSH

I am a Researcher and/or an Engineer working for a project. *As it is becoming more commonplace, and at times required (e.g.,by funders and institutions), to draft Data Management Plans (DMPs), I put my processed research data in trusted Open Science repositories to open my data for potential reuse, allow for citation of my work, and enhance trust in my research. I also want to have an idea of who uses my datasets and how.*
I am a Research software engineer working for a project and/or research infrastructure. *Since I am involved in continuous data collection and handling by means of the software I am maintaining, I am also interested in the tools other researchers use for handling the dataset.*
I am a Manager of a data repository. *I want to understand the use and citation of the research data hosted by my repository, so that I can show qualitative and quantitative figures (e.g., to my funders).*
I am a Data Librarian, a Data Steward or an Open Science Officer. *I support the work of researchers and provide guidance for best practices for research data citation and reuse in their research projects.*
I am a Research Funder*. I want to have a clear view of the "degree of compliance" regarding the DMP submitted by the research project. I want to have a "citation index" of datasets financed to demonstrate the impact of our investment or identify understudied or underfunded research subjects.*
I am a Researcher who is conceiving a research project. *I wish to see what has already been done as I investigate the feasibility of a future project, as a sort of "data bibliography" to understand what research data already relates to the subject or to reuse existing data.*
I am a member of the public who wishes to reuse data. *Either in my work as a journalist, data scientist, or simply an interested citizen, I can find and reuse datasets via proper data citation.*

Recommendations for FAIR Data Citation in the SSH

Societal/Technical Challenge (adapted from FORCE11 principles)	**Recommendation**	**Expected Outcomes**
Importance: Current lack of a "data citation culture," "disappearance" of research material by non publication.	➢ Publish, when applicable, in data journals ➢ Preserve and disseminate to avoid disappearance and minimize costs ➢ Outline an importance of treating research data as individual research outcome complementary to paper ➢ Provide a ready to use "cite as" recommendation in one's published works or datasets	➢ Raising awareness on research data citation ➢ Minimizing costs of research data by fostering their reuse ➢ Endorsing reuse of research data ➢ Improving the use of standards for research data ➢ Increasing visibility and value of research data hidden behind the paper
Credit and Attribution: Creating a dataset (from raw to processed data) in SSH is arduous work carried out by a professional team, and at present these efforts are not sufficiently credited both in terms of intellectual recognition and career recognition.	➢ Use standardized ways of giving credit to authors, contributors, and funders, for instance: at a personal level via ORCID; at project level via GRANTS ID, etc; at institutional level via institutional IDs etc. (e.g.,make those responsible for research data contribution explicit in a paper) ➢ Use standardized licences that correspond to the research data and disciplinary practices (e.g.,Creative Commons if possible, or other specific licences like ODbLas necessary) ➢ Engage with stakeholders to recognize the output	➢ Raising awareness of contributor's work for stakeholders, such as funding agencies publishers, universities and research institutes ➢ Ensuring that all people involved in the creation of research data receives proper credit ➢ Clarifying procedure for creating datasets so that it becomes clear and traceable ➢ Clarifying the legal framework for citation and reuse ➢ Getting recognition of outputs (e.g.,from funding agencies)

Evidence: Proper citation of research data is necessary to understand, trust, and verify the research process: For instance, there is no common understanding or practice around what constitutes evidence and how you reference it.	➢ Encourage the citation of datasets by researchers ➢ Encourage data users to document research questions and investigations referring to the dataset	➢ Reinforcing the importance of research data citation in the SSH ➢ Giving credit to researchers who perform the intellectual work to constitute SSH data ➢ Fostering trust and making the process of creating research data more transparent and understandable ➢ Demonstrating the reproducibility of results, and thus lending credibility, through contextualization of the research process
Unique Identification: Although the use of persistent identifiers is increasing for datasets, these datasets show a great level of variation in terms of volume and internal complexity, much greater than research papers or books standing on the same bookshelf, forming easily comparable units. Therefore, there is a need to refer to the appropriate level of granularity for a research object or a part of an object (e.g.,recording sample, part of an image using IIIF, sentence in a text etc.).	➢ Make research data easier to find and cite with the systematic use of PIDs whatever the technology ➢ Favor using the most discrete citation possible (e.g.,citing a given image of a digitized manuscript, not simply the entire digitization) ➢ For specific data, like dynamic data (e.g.,from social media) try to find the most adequate and persistent way to refer to it (e.g.,permalinks if nothing else is available). In this context, see also "Digital Trace Data"	➢ Enhancing discoverability, identification and accreditation of research data (e.g.,being integrated in Knowledge Graphs such as those operated byFREYA or OpenAire) ➢ Paying attention to granularity when it comes to implementing PID systems or citing PIDs brings scholarly transparency, enhanced reproducibility or reusability and provides a great help in understanding the data architecture as well as scholarly processes
Access: SSH research data are often provided without supplementary documentation that makes them difficult if not impossible to (re)use.	➢ Provide the most comprehensive metadata possible (e.g.,context, version, use controlled vocabularies etc.)	➢ Promoting contextualization and comprehension of the research data being cited, and therefore, its potential limits ➢ Fostering potential reuse of data

	➢ Link research data to other resources that demonstrate data provenance, such as a description of the context of production (e.g, project proposal) software, publications, blog posts etc. Use, as much as possible, standard and accessible formats (e.g.,See NARA Risk Matrix)	
Persistence: Research data should persist beyond the research project itself. Until recently, SSH data was not considered as a crucial product of the research process, more as a tool used to conduct research.	➢ Create and maintain sustainable infrastructures for SSH in order to achieve persistence ➢ Use trusted data repositories with a clear roadmap and good practices that comply with standards (e.g.,long-term preservation or accessibility, etc.) ➢ Train researchers to build a DMP at the very beginning of the project with the support of data stewards and periodically update it during the project lifecycle (i.e. living document) ➢ Support researchers in the execution of their data management strategy ➢ Only research data that is citable is to be preserved	➢ Enhancing discoverability, identification, accreditation and potential reuse of data ➢ Improving the preservation of research data to help justify and off-set the costs of producing it
Specificity and Verifiability: As research data and datasets evolve, it is important to know which version of a dataset one is referring to.	➢ Build a chain of trust at each step of the research process to ensure the authenticity of research data and to facilitate access to successive versions,	➢ Fostering and enhancing specificity, authenticity and verifiability of research data

When citing a part of a dataset, sometimes scholars forget the citation for the whole dataset, and stop with the specific citation.	which can be interesting for SSH research ➢ Use versioning, e.g., from raw data to processed data with a link to the used tool (e.g.,OCR, Visualisation, AI, NLP etc.) ➢ Encourage research projects to cite which version of dataset they are referring ➢ Always give the most complete citation possible, whether that be the smallest atomic part of a citation or the larger project or the ensemble of datasets from which it comes	➢ Building trust in the provenance of research data ➢ Enhancing the re-use and reproducibility of research in SSH
Interoperability and Flexibility: Research data in SSH are very diverse (from geographical data to social media). There is a need to define a common framework of description which doesn't yet exist, that would allow the flexibility to address the different needs and diversity of SSH disciplines and communities while also providing the framework for interoperability.	➢ Use standards for research data and metadata (e.g.,RDF, DCAT, DataCite etc.) and trusted repositories (e.g.,certified, endorsed by communities, subject based etc.) ➢ Use standard communication protocols (e.g.,OAI-PMH) and APIs	➢ Fostering reutilization of research data ➢ Tackling data diversity issues by making them interoperable ➢ Standardizing access to data ➢ Building strong consensus in research data sharing

Annex References

- FAIR Principles, https://www.go-fair.org/fair-principles/
- Data Citation Principles, https://www.force11.org/datacitationprinciples
- Nicolas Larrousse, Daan Broeder, Jan Brase, Cesare Concordia, & Vasso Kalaitzi. (2019). SSHOC D3.2 Inventory of SSH citation practices, and choice for SSHOC citation formats and implementation planning (Version v1.0). Zenodo. https://zenodo.org/record/3595965#.YPlqT-gzY2x
- Erzsébet Tóth-Czifra. (2020). 10 practical tips to fight against the culture of non-citation in the humanities. *DARIAH Open*. https://dariahopen.hypotheses.org/747
- Jennifer Edmond & Erzsébet Tóth-Czifra. (2018). Open Data for Humanists, A Pragmatic Guide. Zenodo. http://doi.org/10.5281/zenodo.2657248
- FAIR Data-Citation for Social Sciences and Humanities, https://www.eosc-hub.eu/events/realising-european-open-science-cloud/fair-data-citation-ssh
- Contributor Roles Taxonomy, NISO, http://credit.niso.org/
- Registry of Research Data Repositories, https://www.re3data.org/
- Creative Commons, https://creativecommons.org/share-your-work/
- Open Data Commons, Legal Tools for Open Data, https://opendatacommons.org/licenses/odbl/
- Digital Trace Data, https://sicss.io/2019/materials/day2-digital-trace-data/what-is-digital-trace-data/What_is_Digital_Trace_Data.html#/
- Digital Preservation Risk Matrix, https://github.com/usnationalarchives/digital-preservation/tree/master/Digital_Preservation_Risk_Matrix

SSHOC
social sciences & humanities open cloud

www.sshopencloud.eu

@SSHOpenCloud

in/company/sshoc

info@sshopencloud.eu