Diary of our initiatory journey on the continent of data citation in SSH

Nicolas Larrousse, Edward J. Gray, Cesare Concordia

Diary of our initiatory journey on the continent of data citation in SSH

The metaphor of a travel journal of an expedition seemed appropriate to us to present this work carried out during the SSHOC[i] project.

The first part was to study this terra incognita by making an inventory of citation practices[ii]. To summarize, we discovered that in the SSH research communities we investigated, practices were seldom standardized and were very diverse, generally producing citations that could not be processed by machines: in other words they were not “actionable”.

This led us to develop a sort of guide necessary to journey through this new, uncharted territory in the form of a set of recommendations[iii] to build citations in SSH. So as not to reinvent the wheel, we based these recommendations on existing principles created by Force11[iv] by adapting them to the specific characteristics of the SSH data. These recommendations were validated by a committee of experts from different backgrounds and structures (RDA participants, CODATA director, OpenAire Engineers etc.) during a round table[v] and in a parallel review process.

Then we decided to analyze the resources available in this new territory, that is, the repositories that are so crucial to be able to cite data. We carried out an analysis of 85 repositories against 7 quality criteria to address the “challenges” described in the recommendations mentioned above:

- **PID** from “Unique Identification & Persistence”
- **Landing page** from “Access”
- **Structured metadata** from “Importance & Credit and Attribution”
- **Cite as** from “Evidence, Specificity & Verifiability”
- **Versioning** from “Specificity and Verifiability”
- **Standardized vocabularies** from “Interoperability and Flexibility”
- **Links to publications** from “Importance”

The results of this survey[vi] are encouraging - even if there is room for improvement, particularly in the use of Persistent Identifiers. Importantly, the presence of a landing page in almost all cases allowed us to build up a test sample made up of a very diverse dataset from those repositories for which we want to build standardized and actionable citations.

In parallel we developed a tool in order to “harvest” the resources found in this new land so as to better understand them and also be able to explain them to others. We developed a prototype composed of three components:

- a harvester which grabs information about a dataset and normalizes it based on the work done by SCHOLIX[vii]
- an API to disseminate the metadata of the citation thereby making it actionable
- a citation viewer for human purposes

For the first iteration to populate this prototype, we used the dataset collected during our survey of repositories and we are going to gradually add more datasets from various sources.

This prototype is primarily designed to implement what we called “actionability” to a citation and provide a ready-to-use citation in various citation formats. Starting from the PID of a dataset, the prototype attempts to aggregate metadata from different sources: the repository of the dataset, the PID Registration Agency and a number of Knowledge Graphs. For instance, while metadata associated with a DOI (Digital Object Identifier) are limited and those provided by a handle are even more scarce, it is possible to get more information from a landing page and thus enrich the citation.
We also used another indirect approach to gather additional information by using a registry of repositories (RE3Data[viii]) which provides, among other things, information on the available APIs available for a specific repository.
Thus, the prototype can give a unified view of information about datasets coming from different sources. For researchers, it thus avoids cumbersome work on how to cite a dataset or get information about its provenance. In return, it makes a researcher aware of the importance of properly documenting a dataset and depositing it in a “good” repository.
The code of the prototype is available on the GitLab instance maintained by ISTI-CNR.
This paper will present in greater detail what we learned at **each step of this expedition** and how a research project can take advantage of a good citation system to enhance the visibility of the output. We will also introduce the potential uses based on the information provided by the prototype such as the possibility of associating a specific tool to process data or the use of this information as a base to build data papers.

Bibliography

1. Blaney, Jonathan. (2012). 'The Problem of Citation in the Digital Humanities'. In: Clare Mills, Michael Pidd and Esther Ward. *Proceedings of the Digital Humanities Congress 2012*. Studies in the Digital Humanities. Sheffield: The Digital Humanities Institute, 2014. Available online at: https://www.dhi.ac.uk/openbook/chapter/dhc2012-blaney

2. Rauber, Andreas, Asmi, Ari, van Uytvanck, Dieter, & Proell, Stefan. (2015). Data Citation of Evolving Data: Recommendations of the Working Group on Data Citation (WGDC). https://doi.org/10.15497/RDA00016

3. Task Group on Data Citation Standards and Practices, 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

[i] https://sshopencloud.eu/
[ii] https://doi.org/10.5281/zenodo.3595965
[iii] https://doi.org/10.5281/zenodo.5361717
[iv] https://doi.org/10.25490/a97f-egyk
[v] https://www.sshopencloud.eu/news/roundtable-experts-data-citation
[vi] https://doi.org/10.5281/zenodo.5603306
[vii] http://www.scholix.org/

[viii] https://www.re3data.org/