The SSH Data Citation Service, A Tool to Explore and Collect Citation Metadata

Cesare Concordia[1](✉), Nicolas Larrousse[2], and Edward Gray[2]

[1] Institute of Information Science and Technology – CNR, Pisa, Italy
cesare.concordia@isti.cnr.it
[2] Huma-Num CNRS, Paris, France
{nicolas.larrousse,edward.gray}@huma-num.fr

Abstract. This paper presents the SSH Data Citation Service (DCS), a software tool that provides functionalities to find, collect and analyse metadata related to digital objects, in particular datasets, referred to in citation strings. Starting from the citation string of a dataset, the DCS aggregates metadata related to the data from different sources: the repository hosting the dataset, PID Registration Agencies and Knowledge Graphs and gives a unified view of information about datasets coming from these sources. The DCS has been designed and developed in the Social Sciences & Humanities Open Cloud (SSHOC) project. It has been used in a project activity as a tool to help investigate approaches adopted for data citation by Social Sciences and Humanities organisations managing data repositories, and as an utility to help data managers to create citation metadata. The paper presents motivations underlying the creation of the tool, the design principles adopted, an overall description of the functionalities of the current release and a summary of ongoing activities.

Keywords: Data citation · Social Sciences and Humanities · Data repositories and archives

1 Introduction

This paper presents a software tool, designed and developed in the Social Sciences & Humanities Open Cloud (SSHOC) project[1], that finds, collects and analyses metadata related to a data citation string. One of the activities carried out in the SSHOC project has been to investigate approaches adopted by organisations and research groups publishing data in the Social Science and Humanities (SSH) domains, to implement standards and recommendations on data citation. The investigation started by making an inventory of citation practices and analysing the approaches followed by main communities in SSH domains. The result of this phase is described in detail in Deliverable 3.2 of the project[2], essentially in the communities investigated, practices were seldom standardised and were very diverse. The second part of the activity has been to define a set

[1] https://sshopencloud.eu.
[2] https://doi.org/10.5281/zenodo.4436736.

G. Silvello et al. (Eds.): TPDL 2022, LNCS 13541, pp. 351–356, 2022.
https://doi.org/10.1007/978-3-031-16802-4_32

of recommendations[3] to build citations for SSH data, based on principles defined by Force11 [7]. These recommendations have been discussed and validated by a committee of experts during several internal events and in a public round table[4]. The final part of this activity has been to analyse 85 repositories identified during the project, to check which of the defined recommendations are implemented by each of them, results of this survey[5] are encouraging - even if there is room for improvement, particularly in the use of Persistent Identifiers (PID). The SSH Data Citation Service (DCS) is a software tool developed during this activity. Starting from the citation string of a dataset, the DCS aggregates related metadata from different sources: the repository hosting the dataset, PID Registration Agencies[6] and a number of Knowledge Graphs. Thus the DCS gives a unified view of metadata related to datasets coming from different sources. It provides some functionalities to analyse the metadata, in particular actionability and interoperability metadata. This paper presents main motivations underlying the creation of the tool, the design principles adopted and an overall description of its functionalities. At the time when this paper is written a prototype of the DCS is published and activities are in progress to release a stable version.

2 The Citation Metadata

The great heterogeneity of scientific data, and the variety of data management systems adopted to publish it, has outlined the importance of creating specific practices and guidelines for 'data citation' [1]. There are several recommendations and best practices that describe the information that a citation string should include, typically the citation should identify the data source, the authors, the publisher, the terms of use of data etc. According to FORCE11 principles [7], data citations should [also] facilitate access to the data themselves and to associated metadata, documentation, code, and other materials, as are necessary for both humans and machines to make informed use of the referenced data. We can say that the role of data citation does not end with the ability to attribute credits, just as important is the ability to enable the reuse of the cited data. However the information needed to reuse data [2, 5], especially those needed by automatic agents, are not present in the citation string (which typically includes information to attribute the data); this information is usually published in metadata records associated with the data. This is due to the changeability and the impermanence of digital data: a dataset could be migrated to new formats and stored in a different management system, data in a repository may be deleted, completely or partly, for instance due to modification in licences. Furthermore, datasets having complex structures could require a reorganisation that may affect this information. There is a need to ensure that a citation remains consistent despite these potential changes: the use of a PID Persistent Identifier (PID) could be a first step but it's not enough. The set of metadata associated with a dataset is defined in the data management and stewardship plan adopted by a publisher; there are good practices and guidelines for creating these plans, however there are several

[3] https://doi.org/10.5281/zenodo.5361717.
[4] https://www.sshopencloud.eu/news/roundtable-experts-data-citation.
[5] https://doi.org/10.5281/zenodo.5603306.
[6] https://pidservices.org/.

different metadata models (in some cases domain specific) and the metadata published may not contain enough information to facilitate the access to the data. The SSH Data Citation Service has been designed to help investigate this specific aspect of the data citation in SSH domains: it enables researchers to discover, collect and analyse the metadata associated with published datasets and it may help to *enrich* the citation with the metadata records collected.

3 The SSH Data Citation Service

The SSH Data Citation Service is a software tool designed and developed to retrieve and analyse metadata related to the digital object referred in a citation string, the collected metadata may be visualised in a web based GUI, stored as JSON objects and possibly processed by software agents. The DCS is designed according to the classical client server architecture: the backend implements the discovering, the management and the persistence of metadata, the client, called Citation Metadata Viewer, shows the metadata and provide actionability functionalities; a REST API implements the interaction protocol between client and server components, and can also be used also as the integration layer with third party systems and agents. The connection of the DCS with an Authentication, Authorization, and Accounting Infrastructure (AAAI) is not yet completely implemented; currently a token authentication mechanism is used. Web based technologies have been used for the implementation: Angular JS framework for the frontend, Java language and technology framework for the backend, the source code of the DCS current release is available.

3.1 Getting the Metadata

The DCS uses the data citation infrastructure [4] to retrieve metadata related to a dataset. The data citation infrastructure is the technological infrastructure that implements referring to data in a unique and persistent manner, it is built upon existing scholar infrastructures and provides functionalities for not just referring to data, but also to making data reusable [3]. Technically speaking it is an heterogeneous infrastructure whose key components are servers resolving identifiers of digital objects and frameworks managing repositories. The citation metadata are published by the components of this infrastructure, however mechanisms and protocols to retrieve this information are very diverse and depend on the components used to publish them. The DCS uses the citation string to find the metadata; it parses the string, extracts the identifier, and retrieve the metadata using different mechanisms:

- **Getting metadata from PID Registration Agencies (RA)**. The PID RAs are crucial components of the citation infrastructure, their role is to provide services that enable organisations to create a Persistent ID for digital objects and to implement the association of the PID with the link of the digital object. Some of the RAs, in particular those managing Digital Object Identifiers (DOI), also provide services for hosting and publishing metadata describing digital objects; these metadata are created by organisations when registering persistent identifiers for their digital objects. The metadata

models provided for datasets usually include many of the information required to access the data. If the identifier of the data is a DOI the DCS try to retrieve metadata using RA API[7] and or content negotiation[8].

- **Getting metadata from landing pages**. The DCS checks if the identifier refers to a web page, web pages linked by identifiers contained in citations are called landing pages. They are human readable documents describing the cited resource, that also provide links and information for accessing the actual data. This information may also be present in the landing page as machine readable metadata; the DCS parses the source code of landing pages to extract the metadata.
- **Getting metadata using repositories API**. Many repositories provide an API to access the data, the DCS in these cases try to extract metadata using the API. To do this it uses information stored in the R3Data registry[9]. The R3Data registry contain information about repositories and, if the repository provides API, this information includes the type of the API provided and the API entry string. The DCS identifies the repository by dereferencing the identifier; if the repository is registered in R3Data and has an API, the DCS uses the API to try to obtain metadata. Currently it is implemented for the OAI-PMH API.
- **Getting metadata from knowledge graphs**. A Knowledge Graph (KG) is a collection of research objects interlinked by semantic relationships. A number of knowledge graphs containing information about SSH datasets exist, among there: the FREYA PID graph[10], the ResearchGraph built by the ResearchGraph Foundation[11], the Research Graph built in OpenAIRE[12]. The KGs do not merely contain bibliographic metadata, they also provide semantic descriptions of scholarly knowledge in the form of actionable statements. The semantic statements and their models may be defined by computational logic-based ontology languages. Knowledge graphs may contain significant information to help implement data citation accessibility and an activity is in progress to implement a module in the DCS to extract metadata from KGs (Fig. 1).

The metadata collected by the DCS is shown to a user by the Citation Metadata Viewer and can be stored locally as a JSON object.

3.2 Data Citation and Machine Actionability

The machine actionable citation metadata is the subset of the metadata that may enable a software agent to automatically identify the structure of the cited data and in some cases process it. The DCS provides a functionality to individuate this kind of information in the collected metadata and to enrich the metadata with information that could be used

[7] Crossref: https://github.com/CrossRef/rest-api-doc, DataCite: https://support.datacite.org/reference/introduction, mEDRA: https://api.medra.org/, EIDR: https://www.eidr.org/technical-documentation/.

[8] https://data.datacite.org/, http://data.medra.org/.

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

[10] https://www.project-freya.eu/en/pid-graph/the-pid-graph.

[11] https://researchgraph.org/.

[12] https://graph.openaire.eu/.

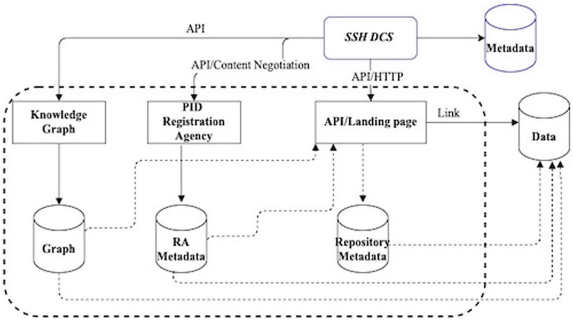

Fig. 1. DCS and citation infrastructure

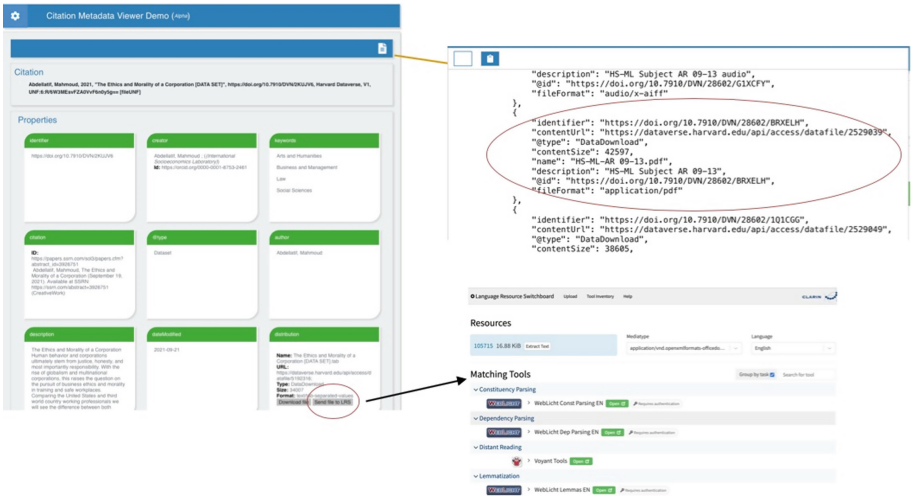

Fig. 2. Citation metadata viewer

by an automatic agent. An example of this functionality is shown in Fig. 2: the metadata associated to the dataset referred by an identifier includes metadata values that can be used for actionability. The DCS recognizes it and the Citation Metadata Viewer displays a button that activates an external application to process the data. In this example the external application is the CLARIN Language Resource Switchboard[13] (LRS), a tool registry that identifies and presents in a GUI a set of tools that can process a resource; by clicking on the button of the Citation Metadata Viewer the data is automatically uploaded on the LRS and eventually processed by the selected tool.

4 Conclusions and Ongoing Activities

The SSH Data Citation Service has been built within the SSHOC project to support investigating the data citation approaches adopted in SSH domains. The DCS provides

[13] https://switchboard.clarin.eu.

functionalities to discover, analyse and render data citation metadata, i.e. the metadata used to describe dataset referred in a citation, in particular the metadata enabling access and reuse of data. It has been designed and developed from scratch and is composed of two main parts: the Citation Metadata Viewer that implements the visualisation logic, and the back end that implements the business logic and the persistence layer. A prototype of DCS has been presented in a number of public events[14], during these events many suggestions and remarks have been collected about its functionalities and possible improvements. The current activity of the DCS development team is mainly focused on integrating an AAAI and on building a stable release of the tool. An investigation is in place to verify the possibility to extend the DCS software modules that implements the machine actionability of citation metadata to interoperate with the technical solutions created inside two interesting SSH projects: an implementation of CMDI-based Signposting [6], and the "Digital Object Gateway"[15] (DOG) project, which adopts principles defined by the FAIR Digital Objects (FDO) community[16]. Furthermore, a SPARQL-based module to interact with Knowledge Graphs is being developed. Additionally, a functionality to partially automate error detection in collected metadata is being developed. The current release of the Citation Metadata Viewer can be accessed and tested[17], the source code is available in a GIT repository[18].

References

1. Castelli, D., Manghi, P., Thanos, C.: A vision towards scientific communication infrastructures. Int. J. Digit. Libr. **13**, 155–169 (2013). https://doi.org/10.1007/s00799-013-0106-7
2. Hourclé, J.: Advancing the practice of data citation: a to-do list. Bull. Am. Soc. Inf. Sci. Technol. **38**(5), 20–22 (2012)
3. Fenner, M., et al.: A data citation roadmap for scholarly data repositories. Sci. Data **6**, 28 (2019). https://doi.org/10.1038/s41597-019-0031-8
4. Groth, P., Cousijn, H., Clark, T.: Carole Goble; FAIR data reuse – the path through data citation. Data Intell. **2**(1–2), 78–86 (2020). https://doi.org/10.1162/dint_a_00030
5. Silvello, G.: Theory and practice of data citation. J. Assoc. Inf. Sci. Technol. **69**(1), 6–20 (2018). https://doi.org/10.1002/asi.23917
6. Arnold, D., Fisseni, B., Trippel, T. Signposts for CLARIN. In: Selected Papers from the CLARIN Annual Conference 2020 (2021). https://doi.org/10.3384/ecp1803
7. Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. In: Martone, M. (ed.) FORCE11, San Diego CA (2014) https://doi.org/10.25490/a97f-egyk

[14] https://www.sshopencloud.eu/events/fair-ssh-data-citation-practical-guide.
[15] https://www.clarin.eu/sites/default/files/20210610-Dieter_Van_Uytvanck-DOG.pdf.
[16] https://fairdo.org.
[17] https://v4e-lab.isti.cnr.it/citview/demo/index.html.
[18] https://gitea-s2i2s.isti.cnr.it/concordia/sshoc-citationservice.