Laboratory Astrophysics: from Observations to Interpretation
Proceedings IAU Symposium No. 350, 2019
F. Salama & H. Linnartz, eds.
doi:10.1017/S174392131900783X

Citation of evolving data in distributed asynchronous infrastructures

C. M. Zwölf, N. Moreau, Y. A. Ba and M. L. Dubernet

LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne University, 5 Place Janssen, 92190 Meudon, France

Abstract. The VAMDC Consortium intended to find a way for users to cite the datasets accessed through the infrastructure. The Research Data Alliance Data citation working group provided the researchers and data centres communities with a recommendation to identify and cite dynamic data. This recommendation perfectly matched the VAMDC needs: the proposed solution relies on a query centric view and the set-up of a Query Store. Data should be stored in a versioned time-stamped manner and accessed through queries. The Query Store we implemented for VAMDC is interlinked with Zenodo. Since Zenodo is indexed in OpenAIRE and since the latter implements Scholix, VAMDC indirectly implements Scholix via its Query Store. The paper outlines the successes and limitations of the above approach.

Keywords. standards, atomic data, molecular data, astronomical data bases: miscellaneous

1. Introduction

Citation is a key element in the production of new knowledge, since it enhances trust, reproducibility and gives visibility to the author. According to the FAIR principles, most of the data should be re-used in derived works and the role of citation is crucial in open-data-driven science. The Research Data Alliance†, through its Data Citation Working Group‡ and its RDA/WDS Scholarly Link Exchange (Scholix) Working Group§ has defined new models for citation in the digital era.

The Virtual Atomic and Molecular Data Centre (VAMDC, Dubernet *et al.* (2016)) became a pilot for the RDA data citation working group as the VAMDC infrastructure is an example of a distributed system with no central management mechanism, where each atomic and molecular heterogeneous database federated by VAMDC is an autonomous node that implements a set of interoperability protocols and standards. About 90% of the VAMDC inter-connected databases handle atomic and molecular data that are used for the interpretation of astronomical spectra and for the modelling in media of many fields of astrophysics. Other application fields of data contained into VAMDC include atmospheric physics, plasmas, fusion, lightning technologies, environmental sciences, health and clinical sciences.

2. Data citation recommendation and VAMDC implementation

The RDA data citation working group has provided the researchers and data centers communities with recommendations to identify and to cite dynamic data (Asmi *et al.* (2016)). The proposed solution relies on a query centric view and the set-up of a *Query Store*. Data should be stored in a versioned time-stamped manner and accessed through

† https://www.rd-alliance.org
‡ https://www.rd-alliance.org/groups/data-citation-wg.html
§ https://www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg

queries. The *Query Store* (Zwölf *et al.* (2019)) stores all the identified and time-stamped queries together with the relevant metadata. It also gives access to the the data as produced when a given query is executed. Within the context of the RDA recommendation the term "query" has to be understood in its wider sense: it stands for any processing mechanism used to extract data from a computer-based system.

The *Query Store* has been coupled to the VAMDC Portal (https://portal.vamdc.eu/) which provides a seamless access to the VAMDC inter-connected databases. A description of usage of the VAMDC portal in connection to the query store can be found in Moreau *et al.* (2018). In brief, each query on the VAMDC Portal provides access to landing pages identified with persistent identifiers. Those landing pages store the persistent identifier associated with the query, the query itself, the name and version of the node answering the query, the dataset produced by the node while processing the query, together with the bibliographic references extracted from the dataset. As all queries are stored in the *Query Store* (https://cite.vamdc.org) for a period of time, users can find the landing page at a later stage using the persistent identifier that they have kept. In addition the *Query Store* is coupled to Zenodo that allows to obtain DOI for the landing pages. When some data extracted from VAMDC are cited (in papers and/or other datasets) through the DOI obtained by the couple (Query Store/Zenodo), the authors of the works referenced by the VAMDC data receive credit automatically (Moreau *et al.* (2018)).

3. Conclusion and further work

Basically VAMDC has solved a number of issues concerning the data citation of atomic and molecular data retrieved from databases : a reference is associated to every data and it is now possible to save the queried data and references for given datasets. We describe in Moreau *et al.* (2018) the possible ways of using the *Query Store* with the VAMDC portal.

Nevertheless we have identified a drawback mainly linked to the query system of VAMDC, and not the *Query Store* concept. The VAMDC query system requests scientific data such as species, processes, wavelength for example, but does not allow a selection based on bibliographical information. As an example we cannot currently request data related to a single reference. There is still some work to be done in order to reach a useful point for the users of databases. In addition we are studying the data citation issue related to data published in papers and we think that the *Query Store* concept linked to the standard format of VAMDC-XSAMS (http://www.vamdc.eu/standards) could bring some advances.

The commission B5 of the IAU could be the right place to study those issues and make recommendations.

References

Asmi, A., Rauber, A., Pröll, S., & van Uytvanck, D. 2016, Treatise in Geochemistry 1 (Oxford and San Diego: Elsevier), p. 17 *EGU General Assembly Conference Abstracts*, vol. 18, p. EPSC2016–7456

Dubernet, M.-L., Antony, B., Ba, Y.-A., Babikov, Y., Bartschat, K., Boudon, V., *et al.* 2016, *J. Phys. B: At. Mol. Opt. Phys.*, 49, 074003

Moreau, N., Zwölf, C. M., Ba, Y. A., Richard, C., Boudon, V., Dubernet, M. L. 2018, *Galaxies*, 6, 105

Zwölf, C. M., Moreau, N., Ba, Y. A., Dubernet, M.-L. 2019, *Data Science Journal*, 18, 4

https://doi.org/10.1017/S174392131900783X Published online by Cambridge University Press