Improving Citation of evolving data in distributed asynchronous infrastructures

Implementing RDA Recommendations Data Citation and Scholix in VADMC for molecular and atomic data.

RESEARCH DATA ALLIANCE

Adopting RDA outputs Data Citation and Scholix in the Virtual Atomic and Molecular Data Centre (VAMDC)

The VAMDC Consortium intended to find a way for users to cite the datasets accessed through the infrastructure. The RDA 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.

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: the role of Citation is crucial in open-data-driven science. The RDA has successfully defined new models for citation in the digital era.

Dr. Carlo Maria Zwölf
Virtual Atomic and Molecular Data Centre (VAMDC) executive office, Observatoire de Paris

The Challenge

Using VAMDC, scientists can easily discover atomic and molecular resources, and access their data in a unique and practical way. The adoption of VAMDC by a large community revealed a new set of challenging issues. The VAMDC data are dynamic: a database may evolve over time. It may happen that some of these database evolutions are not systematically reported through new publications. Assuming that a scientist extracts data from VAMDC, and wishes to use this data in order to perform research that will be published into a paper: how can the researcher cite this data? Since the database content may evolve, for the consistency of the scientific publication, the citation should refer to data well defined in space and time. The citations should contain pointers to the authors who originally produced those data. And, to guarantee the reproducibility of the scientific process described in the paper the data should be referenced.

What is your stakeholder community?

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.

How did you identify RDA as a potential support?

In the VAMDC consortium, we started our reflection about data-citation around 2014. We thought this problem was not specific to our case and decided to join the RDA to address and discuss our data-citation issues in a wider community of data practitioners. The first plenary we attended was the RDA Plenary 3rd in Dublin, on March 2014. We discovered there the Data Citation Working Group and started being active members. VAMDC then became an early pilot for this working group.

Find out more at:
www.rd-alliance.org/recommendations-outputs

Visit rd-alliance.org or write us at enquiries@rd-alliance.org

The RDA output adopted

1. RDA Data Citation recommendation: provides the VAMDC users with the capability to uniquely refer-to and cite a given datum extracted from VAMDC. This increases the FAIR-ness of our data.

2. Scholix: motivates data producers to share their data through the VAMDC infrastructure rather than through personal websites, reducing fragmentation.

Implementation

Since the RDA Plenary 3rd in 2014, VAMDC became a pilot for the Data-Citation WG and the recommendation that was being defined: the prototype of that recommendation came for standalone databases or data warehouses. From an architectural point of view, the VAMDC infrastructure may be seen as a distributed system with no central management mechanism: each database federated by VAMDC is an autonomous node implementing a set of interoperability protocols and standards. Our work in the WG consisted in checking that on a conceptual level the recommendation contained no issue, with respect to our architecturally distributed set, preventing the application of the recommendation to VAMDC. In 2017 with a RDA EU 3.0 grant, we were able to pass from the conceptual to the operational level: we implemented the RDA Data Citation recommendation into a software solution designed to work in the case of asynchronous, decentralized and distributed services.
The software solution we implemented had to deal with a lot of constraints. Any evolution of the infrastructure automatically impacts all the connected databases. As a consequence, the majority of the VAMDC Consortium members must validate any technological evolution of the infrastructure.
Any adopted solution must lessen the effects on the existing infrastructure and have minimal implementing costs for each autonomous database federated by VAMDC. We deployed this software on the VAMDC infrastructure, now an operational Query Store accessible at: https://cite.vamdc.eu.

Lessons learned

We succeeded in implementing the new RDA-data citation paradigms (Data Citation and Scholix) on the VAMDC distributed infrastructure. This removed the technical barriers preventing automatic data citation and delegation of bibliographic credits for VAMDC extracted data.
The success of a technical solution does not only depend on its intrinsic qualities, but also on its level of adoption and acceptance by the wider community of final users: we are actually focusing on increasing the impact of the aforementioned citation service through community awareness-raising and training around these new tools.

VAMDC consortium

Virtual Atomic and Molecular Data Centre

The **VAMDC Consortium** is

a consortium of Institutes and Research Institutions that share a common technical and political framework for the distribution and curation of atomic and molecular data.
The VAMDC Consortium technical framework relies on the use of the e-science VAMDC infrastructure that provides the international research community with access to a broad range of atomic and molecular (A&M) data compiled within a set of A&M databases accessible through the provision of a single portal. Furthermore, VAMDC aims to provide A&M data providers and compilers with a large dissemination platform for their work.

Dr. **Carlo Maria** Zwölf,
Nicolas Moreau,
Yaye-Awa Ba,
Virtual Atomic and Molecular Data Centre executive office,
Observatoire de Paris, PSL Research University, CNRS,
Sorbonne University, France
Carlo-maria.zwolf@obspm.fr;
https://orcid.org/0000-0002-5762-6747