

# Final Report

## Query Store for the VAMDC e- infrastructure

---

24/11/2017

Carlo Maria Zwölf, carlo-maria.zwolf@obspm.fr

### Executive Summary

The Virtual Atomic and Molecular Data Center (VAMDC) Consortium is an international consortium of Institutes and Research Institutions that share a common political and technical framework (the VAMDC e-infrastructure) for the distribution and curation of atomic and molecular data.

The VAMDC e-infrastructure federates in an interoperable way ~30 heterogeneous databases.

Since the infrastructure is designed for scientific usage, we pay special attention to problems linked with Data Citation: we have started on Spring 2014 a collaboration with the Research Data Alliance and its Data Citation Working Group (DCWG). VAMDC became one of the early use-cases.

In this collaboration project, we implemented the RDA-DCWG recommendations by providing the VAMDC infrastructure with an operational Query Store.

The VAMDC Query Store has been designed and developed by the VAMDC team (hosted by the Paris Observatory) managing the evolutions and the maintenance of the core components of the e-infrastructure.

The produced software has been used for running the VAMDC Query Store and is published on a Github public repository, together with a full documentation.

It is designed for being easily adapted and re-used by the community.

### Objectives

The goal of this collaboration project was to build a set of methods, strategies and software for providing the VAMDC research e-Infrastructure with a Query Store. This component is the core part of the recommendation of the RDA Data Citation Working Group.

The VAMDC is a worldwide-distributed e-infrastructure with no central management system. The proposed solution must cope with this technical complexity: be as generic as possible and minimally invasive for the existing infrastructure components. One of the consequences of these requirements is that the software element we developed is flexible and may be easily re-used by other data-providers from different scientific communities for building their own Query Store.

## Initial State

The RDA Dynamic Data Citation Recommendation originally comes for standalone databases and warehouses. Nobody tried to implement it on a distributed architecture.

In a preliminary phase, from 2014 to 2016, we collaborated in the RDA context with the chairs of the Data Citation Working Group (VAMDC was one of the preliminary use-cases) and concluded that, from the conceptual point of view, the recommendation could apply with no contradictions in the case of VAMDC. From a technical point of view however, many open and complex problems remained.

These problems were faced and solved in the collaboration project.

## Project Outcomes

To build the VAMDC Query Store, we developed a software layer composed of different modules:

- The server module
- The html Graphic User Interface
- The SQL Query comparator

We deployed this set of software in our infrastructure to obtain a working service: the VAMDC Query Store service is available at the url <http://cite.vamdc.eu>

## The Server Module

This module is the core of the proposed solution and is composed by the set of the Web services setting up the Query Store.

- The public repository for this module is: <https://github.com/VAMDC/QueryStore>
- The Implementing Note is available at <https://github.com/VAMDC/QueryStore/blob/master/documentation/ImplementingNote.pdf>
- The deployment instructions are available at <https://github.com/VAMDC/QueryStore/blob/master/documentation/DeploymentInstructions.pdf>

## The HTML Graphic User Interface

This module is a web interface for the web services implemented by the Server Module. The repository for this module is:

<https://github.com/VAMDC/QueryStoreWeb>. Deployment instructions and documentation are in the Github project readme file.

### The SQL Query Comparator

The VAMDC query Language is a sub-set of SQL. This library reduces VAMDC queries to a standard-normal form and is used to state if two different VAMDC queries are semantically identical. The Server Module uses this library to compute the query uniqueness. The repository for this module is <https://github.com/VAMDC/VamdcSqlRequestComparator>.

## Dissemination Activities / Publications

| # | Event  | Date           | Publication  |
|---|--|----------------|--|
| 1 | RDA Webinar  | March 2017     | Implementing the RDA data citation recommendations in the distributed Infrastructure of the Virtual Atomic and Molecular Data Centre |
| 2 | ASTERICS DADI Technology forum 3                                   | March 2017     | Provenance and Reproducibility. Feedbacks from the VAMDC experience  |
| 3 | IVOA Interoperability meeting                                      | May 2017       | Implementing the RDA data citation recommendations in the distributed Infrastructure of the Virtual Atomic and Molecular Data Centre |
| 4 | RDA Plenary 9 Barcelona  | April 2017     | Implementing the RDA data citation recommendations in the distributed Infrastructure of the Virtual Atomic and Molecular Data Centre |
| 5 | RDA Plenary 10 Montreal  | September 2017 | VAMDC Query Store Usage feedbacks  |
| 6 | 7 <sup>th</sup> ESEP meeting – Databases and Virtual Observatories | November 2017  | The Virtual Atomic and Molecular Data Centre   |

## Summary & Conclusions

We met the project objectives by providing the distributed VAMDC infrastructure with an operational Query Store. The software we implemented

(and the underlying solutions) is flexible and generic in order to be re-used and easily adapted by other data-providers from different scientific communities to build their own Query Store.