

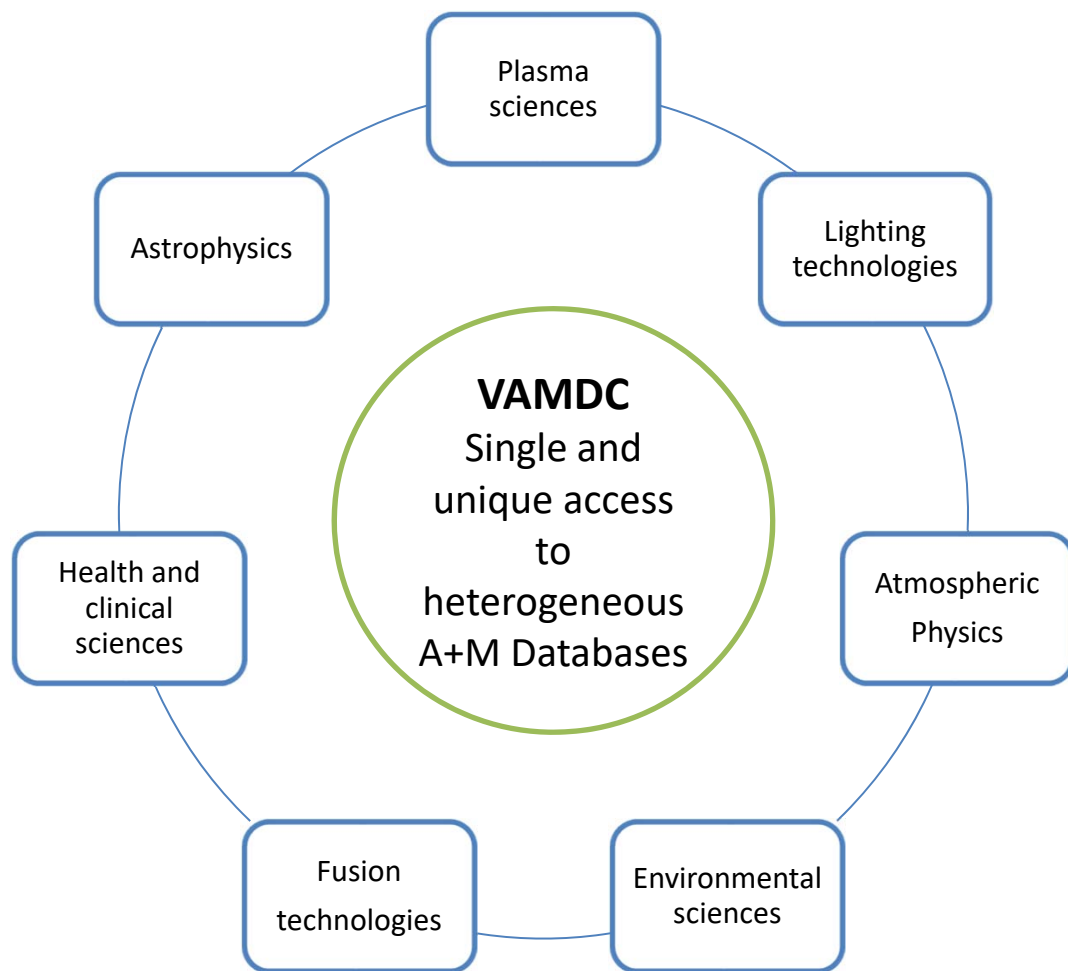
Data Curation & Preservation Shanghai Interop – May 2017

Implementing the RDA data citation recommendations in the distributed Infrastructure of the Virtual Atomic and Molecular Data Centre

C.M. Zwölf, N. Moreau and VAMDC consortium



The Virtual Atomic and Molecular Data Centre



- Federates ~30 heterogeneous databases
<http://portal.vamdc.org/>
- The “V” of VAMDC stands for Virtual in the sense that the e-infrastructure does not contain data. The infrastructure is a wrapping for exposing in a unified way a set of heterogeneous databases.
- The consortium is politically organized around a Memorandum of understanding (15 international members have signed the MoU, 1 November 2014)
- High quality scientific data come from different Physical/Chemical Communities
- Provides data producers with a large dissemination platform
- Remove bottleneck between data-producers and wide body of users

The VAMDC infrastructure architecture

VAMDC:

- Is a distributed e-infrastructure
- Has no central management system
- The federated databases are exposed as TAP services

We provide tools and libraries (called *node-software*) for transforming an existing autonomous database into a TAP service federated to VAMDC

- A federated TAP/Database service is called a *VAMDC-Node*
- *VAMDC-Nodes* are listed into registries (we use the *AstroGrid* one)

The Research Data Alliance and the Data Citation WG

Data Citation WG



i Group details

Status: Recognised & Endorsed

Chair(s): Andreas Rauber, Ari Asmi, Dieter van Uytvanck

Case Statement: [Download](#)

The RDA Working Group on Data Citation (WG-DC) aims to bring together a group of experts to discuss the issues, requirements, advantages and shortcomings of existing approaches for efficiently citing subsets of data. The WG-DC focuses on a narrow field where we can contribute significantly and provide prototypes and reference implementations.

Goals of this WG are to create identification mechanisms that:

- allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner
- allows us to cite and retrieve that data as it existed at a certain point in time, whether the database is static or highly dynamic
- is stable across different technologies and technological changes

Solution: The WG recommends solving this challenge by:

- ensuring that data is stored in a versioned and timestamped manner.
- identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store.

- The RDA recommendations comes from standalone databases or warehouse.
- VAMDC is a distributed infrastructure, with no central management system.

Highlighting the main issues

How to build a Query Store in our distributed infrastructure?

- The solution belongs to a space with lot of constraints
 - Any choice will impact each of the ~30 databases federated by VAMDC.
 - Any technological change of the infrastructure must be validated by the majority of the members

The solution must cause least effects on the existing infrastructure and have minimal implementing cost for the database owners.

- This constraint suggest to fit the solution into the standard wrapping layer transforming an autonomous Database into a VAMDC node.

Sketching the solution strategy

Implementation is an overlay to the standard / output layer, thus independent from any specific data-node

Tagging versions of data

Two layers
mechanisms

1 → **Fine grained granularity:**
Evolution of XSAMS output standard for tracking data modifications*

2 → **Coarse grained granularity:**
At each data modification to a given data node, the version of the Data-Node changes

With the **second mechanism** we know that something changed : in other words, we know that the result of an identical query may be different from one version to the other. The detail of which data changed is accessible using the **first mechanisms**.

Query Store

Is built over the versioning of Data (the coarse-grained mechanism)

Is plugged over the existing VAMDC data-extraction mechanisms.

Due to the distributed VAMDC architecture, the Query Store may be seen as a smart log-service.

* <http://dx.doi.org/10.1016/j.jms.2016.04.009>
arxiv version at <https://arxiv.org/abs/1606.00405>

Let us focus on the query store:

The difficulties we had to cope with:

- Handle a query store in a distributed environment (RDA did not design it for these configurations).
- Integrate the query store with the existing VAMDC infrastructure.

The implementation of the query store is the goal of a joint collaboration between VAMDC and RDA-Europe 3.

- Development started during spring 2016.
- Final product released during 2017.

Collaboration with Elsevier for embedding the VAMDC query store into the pages displaying the digital version of papers.

Designing technical solution for

- Paper / data linking at the paper submission (for authors)
- Paper / data linking at the paper display (for readers)

Further Information

Technical details for data versioning:

New model for datasets citation and extraction reproducibility in VAMDC,

C.M. Zwölf, N. Moreau, M.-L. Dubernet,

In press *J. Mol. Spectrosc.* (2016), <http://dx.doi.org/10.1016/j.jms.2016.04.009>

Arxiv version: <https://arxiv.org/abs/1606.00405>

This mechanism should be included in the follow 2017 version of the standards.

Details about the implemented Query Store:

Video from the last week webinar: <https://youtu.be/OLe-qcqCcCw>

- Information
 - On the software architecture adopted for implementing the recommendation
 - On the implementing tricks

Strengths of the Query Store:

- The QS usage is transparent for users (complexity is hidden).
- Live monitoring of all the queries and users of the VAMDC e-infrastructure
 - Data providers may measure their impact and have detailed statistics of usage.
- It will be easy for authors to cite data coming from VAMDC. Credit to producers will be automatic.

Minimal impact
for federated
database owners
for dealing with
the Query Store

Database owners just need to install the latest
version of the VAMDC node-software

Data providers has to fill a “version” field (~ a simple string),
which is just the version label.

- When the database is modified and/or data node software changes, the version label should evolve.

Concluding Remarks :

- We provided the VAMDC infrastructure with a working Query Store
- **The Query Store works natively for all the TAP services !!**
- The concept adopted and the implemented code are quite generic and both can be adapted to other use-cases:
 - If it worked in our complex distributed case, it may work in many contexts
 - The cost for adapting an existing service/database to a VAMDC-type Query Store is minimal
 - All the complexity is handled and masked in our generic software.

Contact : carlo-maria.zwolf AT obspm.fr

Live demo link: <https://youtu.be/kDDWFpi22cU>