Victoria Interop, May 2018

Query Store for the VAMDC infrastructure

C.M. Zwölf, N.Moreau, Y.A. Ba

Paris Observatory

The Research Data Alliance and the Data Citation WG

About

Research Data Alliance

The Research Data Alliance (RDA) builds the social and technical bridges that enable open sharing of data.

The RDA vision is researchers and innovators openly sharing data across technologies, disciplines, and countries to address the grand challenges of society.

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 agroup 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.

Recommendation is to store all the queries (with their metadata) into a Query Store (QS).

The difficulty we have to cope with
- Handle a QS in the VAMDC distributed environment (VAMDC is a set of independent TAP services)
- Integrate the QS with the existing VAMDC components

QS in a nutshell:

Implementation of the RDA recommendation for the VAMDC infrastructure was done as a part of a sub-contract funded by RDA-EU3 project.
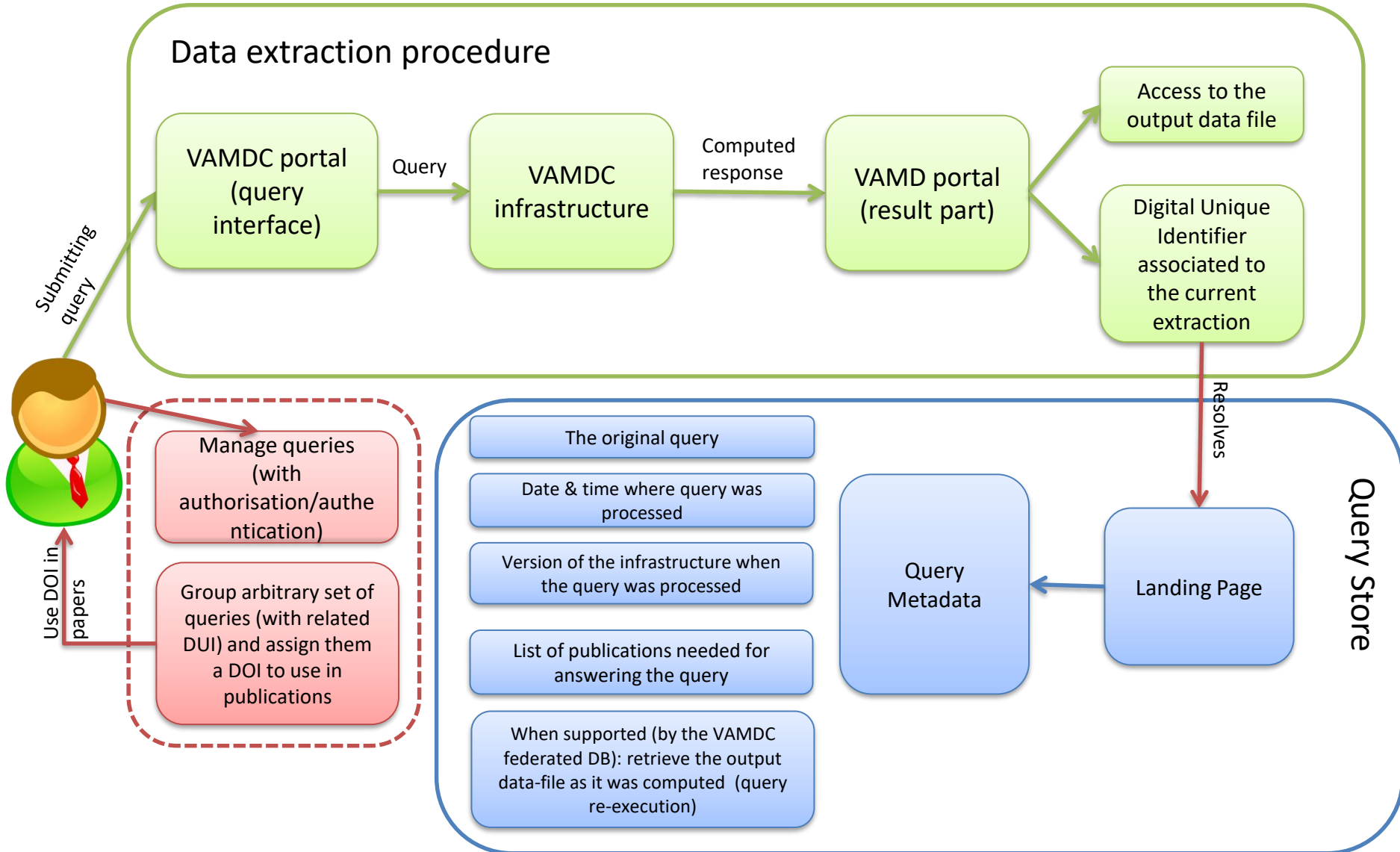
- Made during the year 2017.
- The full source code + documentation is published at https://github.com/VAMDC/QueryStore (CC4ByNDNC license)
- Deployment in production for all the nodes is in progress.
- Technical presentation of the architecture and functioning
  - https://youtu.be/OLe-qcqCcCw

- The Query Store is a central real-time repository for all the queries served by any VAMDC TAP service
  - Independently by the used client for querying the infrastructure
  - Queries are stored together with their metadata:
    - Production/extraction context and timestamp + references + link to the generated XSAMS file + Unique Identifier & DOI
  - Provides live monitoring of the entire infrastructure
    - Data providers may measure their impact and have detailed statistics of usage
  - Increases the quality of the data
    - All the produced XSAMS are parsed for extracting references & validated

Advantage of getting a DOI:

- Demo of new DOI feature: https://youtu.be/CgC-KxOM_8g
- Easy to cite a dataset from its DOI (as currently done for papers)

- XSAMS files are long-lasting and safely stored on the Zenodo Repository.

  - Zenodo is indexed in OpenAire and linked with Scholix (tools used by publishers for computing bibliometrics and impact factors).

  - An author/paper cited by the data-set will get credits automatically when the dataset is cited (using the DOI) into a paper.

    - **Strong marketing argument: Put your data in VAMDC. You will get automatically credits each time your data is cited!**
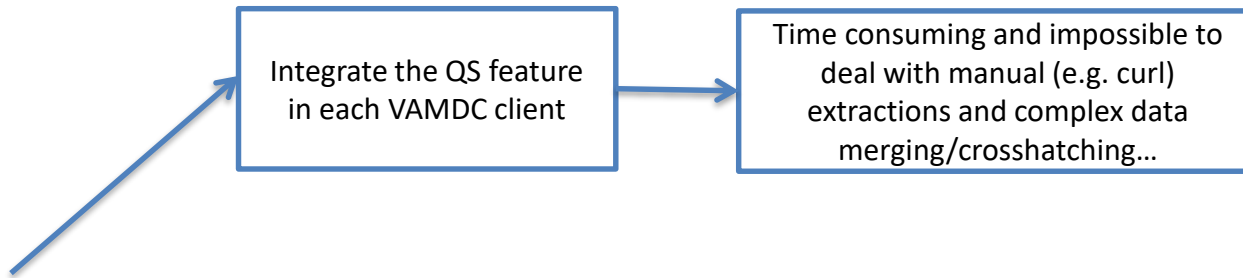
Advantage of getting a DOI:

Data extraction procedure

VAMDC portal (query interface)

→ Query →

VAMDC infrastructure

→ Computed response →

VAMD portal (result part)

Access to the output data file

Digital Unique Identifier associated to the current extraction

Submitting query

Use DOI in papers

Manage queries (with authorisation/authentication)

Group arbitrary set of queries (with related DUI) and assign them a DOI to use in publications

Query Store

Resolves

The original query

Date & time where query was processed

Version of the infrastructure when the query was processed

List of publications needed for answering the query

When supported (by the VAMDC federated DB): retrieve the output data-file as it was computed (query re-execution)

Query Metadata
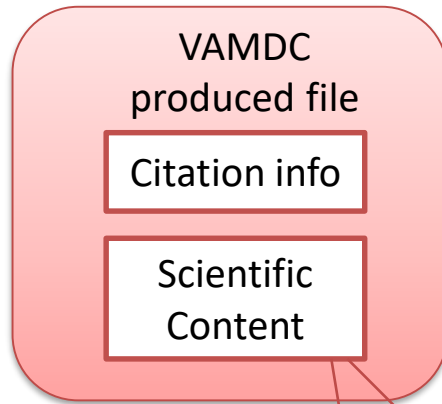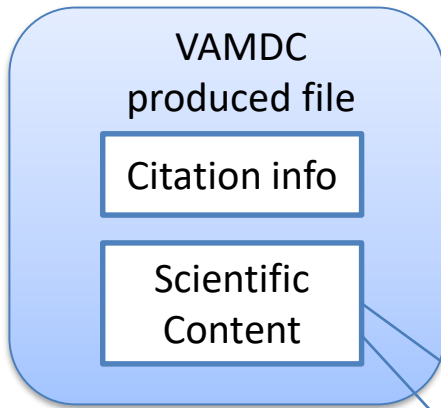
Landing Page

Concluding remarks:

- The current workflow shown how to remove all the technical barriers linked with the automatic data-citation & delegation of credits.

  - Is this workflow suitable for scientists in their everyday activities?

  - How science-tools should evolve for integrating/adopting the QS citation features?

  - What workflow should be optimal for cooperating with authors?

Integrate the QS feature in each VAMDC client → Time consuming and impossible to deal with manual (e.g. curl) extractions and complex data merging/crosshatching…

VAMDC produced file
- Citation info
- Scientific Content

VAMDC produced file
- Citation info
- Scientific Content

VAMDC produced file
- Citation info
- Scientific Content

E.g. production of synthetic spectra.

How to automatize the generation of the citation snippet from these data for this spectrum? How to build the provenance workflow?

Synthetic Spectrum of an Earth-like Planet

O_3 CO_2

CH_4

N_2O

H_2O
CO_2

H_2O

O_2

H_2O

H_2O

CO_2

Flux Density (Jy)

Wavelength (μm)
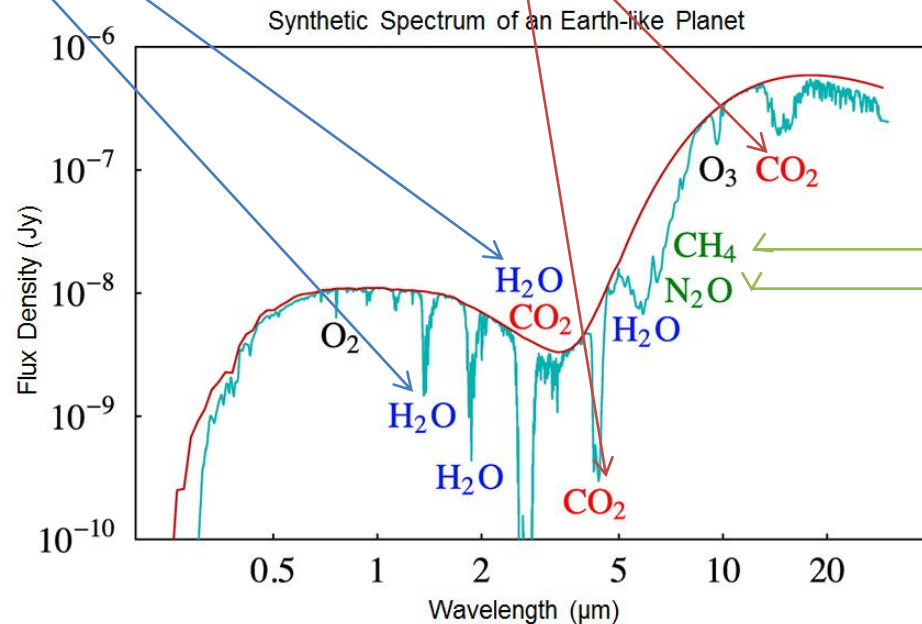
Concluding remarks:

- The current workflow shown how to remove all the technical barriers linked with the automatic data-citation & delegation of credits.

    - Is this workflow suitable for scientists in their everyday activities?

    - How science-tools should evolve for integrating/adopting the QS citation features?

    - What workflow should be optimal for cooperating with authors?

| Integrate the QS feature in each VAMDC client | → | Time consuming and impossible to deal with manual (e.g. curl) extractions and complex data merging/crosshatching… |

| Build a new approach on the existing | → | Promote a new behaviour: users will take trace of how they use data. Once a datum is identified for being published → Ad hoc extraction using the portal | → | Citation workflow will work and we will get a new role and visibility for the portal. |