From RDA Data Citation Recommendations to new paradigms for citing data from VAMDC

C.M. Zwölf and VAMDC consortium

ASOV, Paris – 15/03/2016

VAMDC consortium

A never ending story…

Our common goal and challenge

Produce high quality scientific data

Share them adopting international standards and protocols

Maintain the sharing service(s)

A never ending story…

Our common goal and challenge

Produce high quality scientific data

Share them adopting international standards and protocols

Maintain the sharing service(s)

?? Once achieved, is it the end of the work ??

No!!

A never ending story…

Our common goal and challenge

Produce high quality scientific data

Share them adopting international standards and protocols

Maintain the sharing service(s)

?? Once achieved, is it the end of the work ??

No!!

Our service(s) should be authoritative and acknowledged by the community

Our service(s) must be cited when used in papers

The path of systematic data citation is narrow

The use of a given service is proportional to the ease of access and to its integration in advanced tools (e.g. Topcat) hiding the complexity.

The easy of use and the integration into tools make user forgot he/she is using the service (e.g. -I'm just using Topcat-).
→No citation!

The path of systematic data citation is narrow

The use of a given service is proportional to the ease of access and to its integration in advanced tools (e.g. Topcat) hiding the complexity.

The easy of use and the integration into tools make user forgot he/she is using the service (e.g. -I'm just using Topcat-).
→No citation!

How to deal with these two opposite trends?

The path of systematic data citation is narrow

The use of a given service is proportional to the ease of access and to its integration in advanced tools (e.g. Topcat) hiding the complexity.

The easy of use and the integration into tools make user forgot he/she is using the service (e.g. -I'm just using Topcat-).
→No citation!

How to deal with these two opposite trends?

Is there a way for simplifying, automatizing, enforcing the data-citation?

We have started working on these themes, by joining the **RDA Data Citation Working Group**. VAMDC has become one of the use-case.

The Research Data Alliance and the Data Citation WG

Data Citation WG

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.

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.

Let us implement the recommendation!!

The problem is more **anthropological** than technical…

Tagging and versioning data

What does it really mean *data citation?*

Let us implement the recommendation!!

The problem is more **anthropological** than technical…

Tagging and versioning data

We see technically how to do that

Ok, but What is the data granularity for tagging?

Naturally it is the dataset (A+M data have no meaning outside this given context)

But each data provider differently define what a dataset is.

What does it really mean *data citation?*

Let us implement the recommendation!!

The problem is more **anthropological** than technical…

Tagging and versioning data

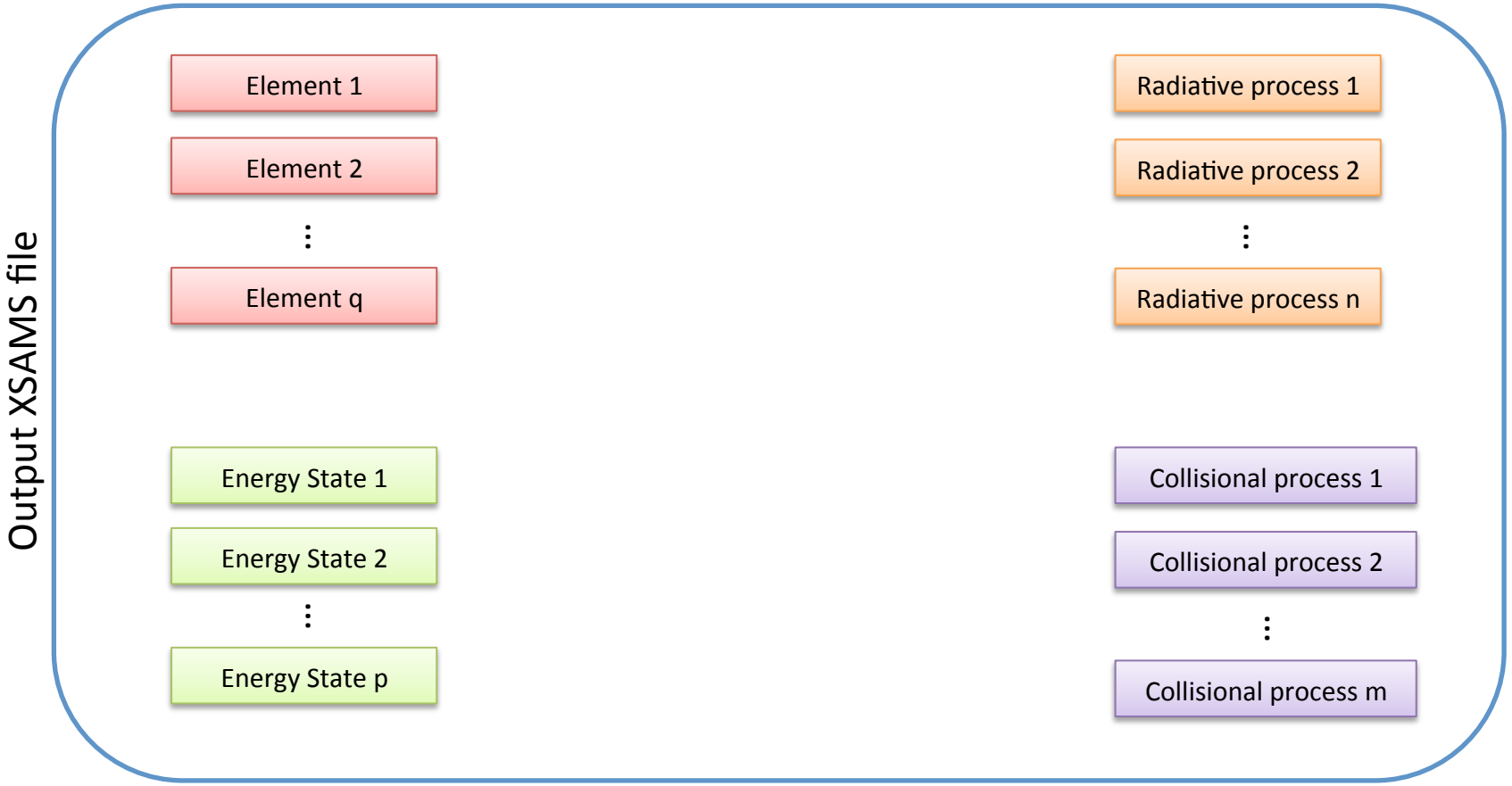
- We see technically how to do that
- Ok, but What is the data granularity for tagging?
- Naturally it is the dataset (A+M data have no meaning outside this given context)
- But each data provider differently define what a dataset is.

What does it really mean *data citation?*

- Everyone knows what it is!
- Yes, but everyone has its own definition
- **RDA** →cite databases record or output files. (an extracted data file may have an H-factor)
- **VAMDC** →cite all the papers used for compiling the content of a given output file.

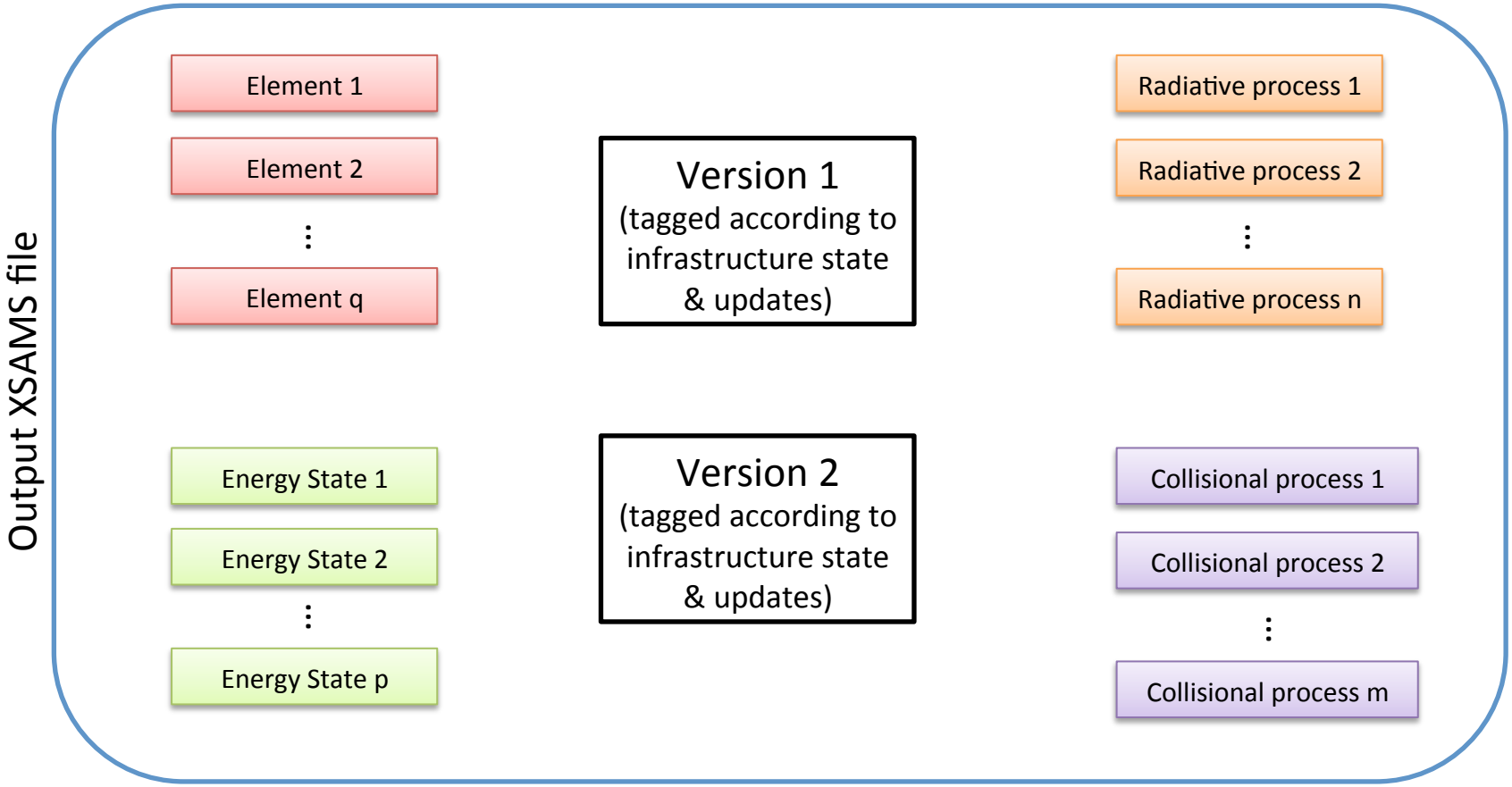
Let us focus on data tagging/versioning issue:

We adopted a change of paradigms (weak structuration):



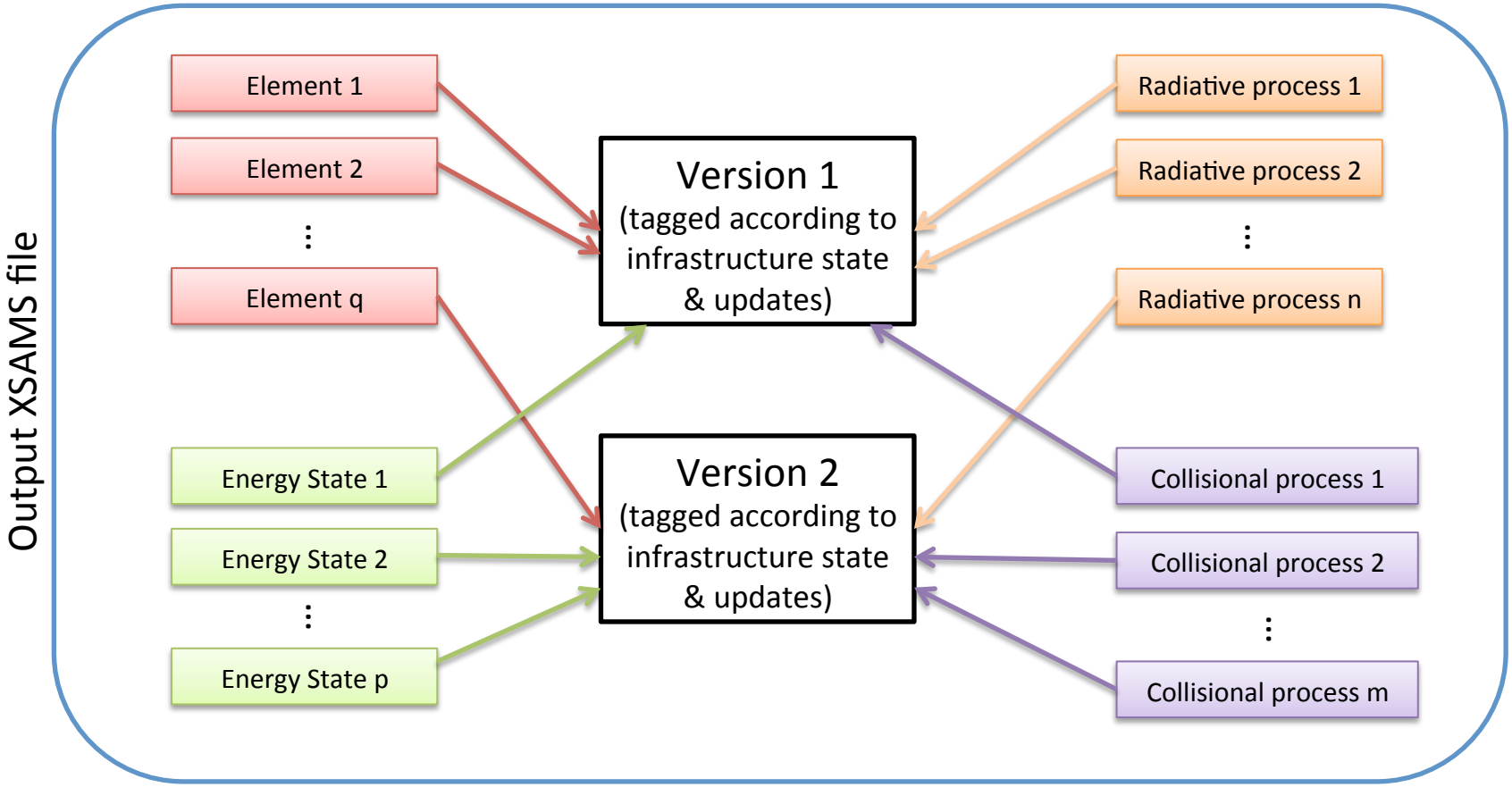
Let us focus on data tagging/versioning issue:

We adopted a change of paradigms:



Let us focus on data tagging/versioning issue:

We adopted a change of paradigms:



Let us focus on data tagging/versioning issue:

We adopted a change of paradigms:

This approach has several advantages:

- It solves the data tagging granularity problem
- It is independent from what is considered a dataset
- The new files are compliant with old libraries & processing programs
  - We add a new feature, an overlay to the existing structure
  - We induce a structuration, without changing the structure (weak structuration)

Let us focus on data tagging/versioning issue:

We adopted a change of paradigms:

This approach has several advantages:

- It solves the data tagging granularity problem
- It is independent from what is considered a dataset
- The new files are compliant with old libraries & processing programs
  - We add a new feature, an overlay to the existing structure
  - We induce a structuration, without changing the structure (weak structuration)

Technical details described in
New paradigm for datasets citation and extraction reproducibility in VAMDC,
C.M. Zwölf, N. Moreau, M.-L. Dubernet,
Accepted into *Journal of Molecular Spectroscopy – Special issue: New visions of Spectroscopic Databases*

Let us focus on the query store:

The difficulty we have to cope with:

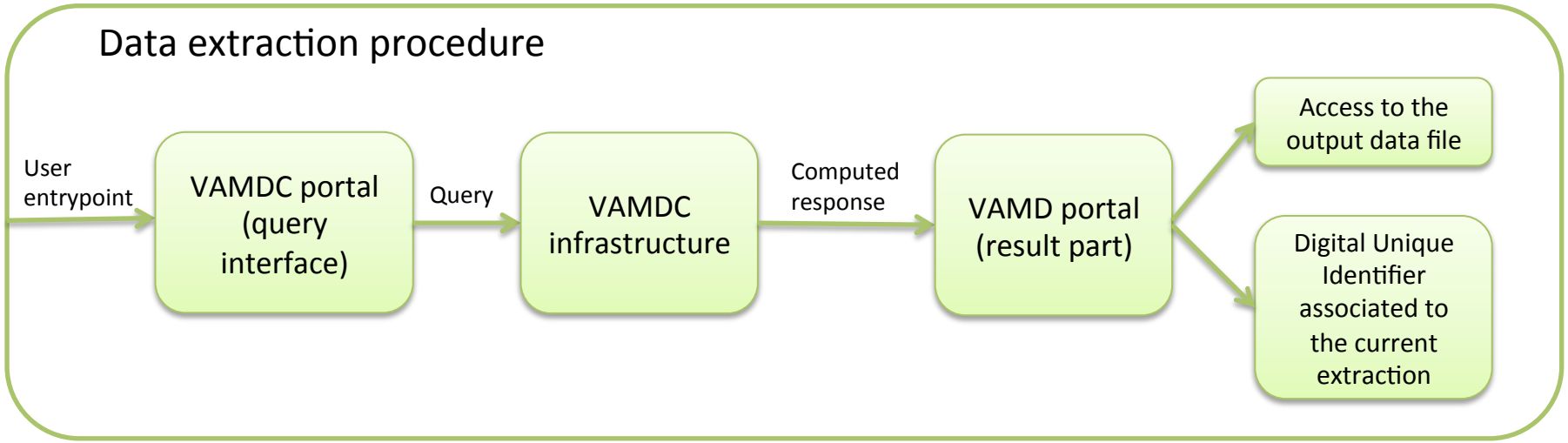
- Handle a query store in a distributed environment (RDA did not design it for these configurations).
- Integrate the query store with the existing VAMDC infrastructure.

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

- Development will start during spring 2016.
- Final product released during 2017.

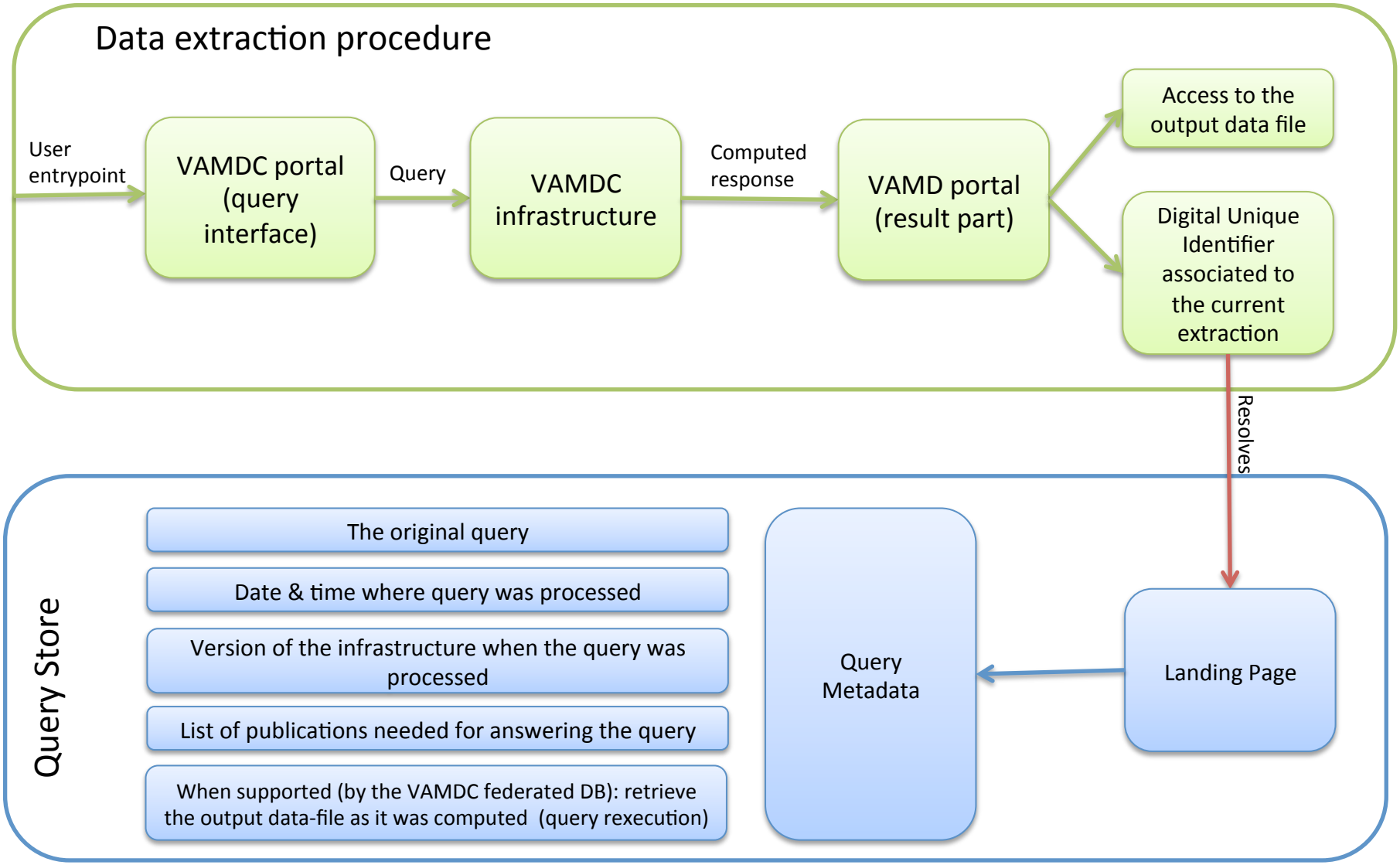
Let us focus on the query store:

Sketching the functioning:



Let us focus on the query store:

Sketching the functioning:



Final remarks:

- Our aims:
  - Provide the VAMDC infrastructure with an operational query store
  - Share our experience with other data-providers
    - Provide data-providers with a set of *libraries/tools/methods* for an easy implementation of a query store.
    - We will try to build a generic query store (i.e. using generic software blocks)