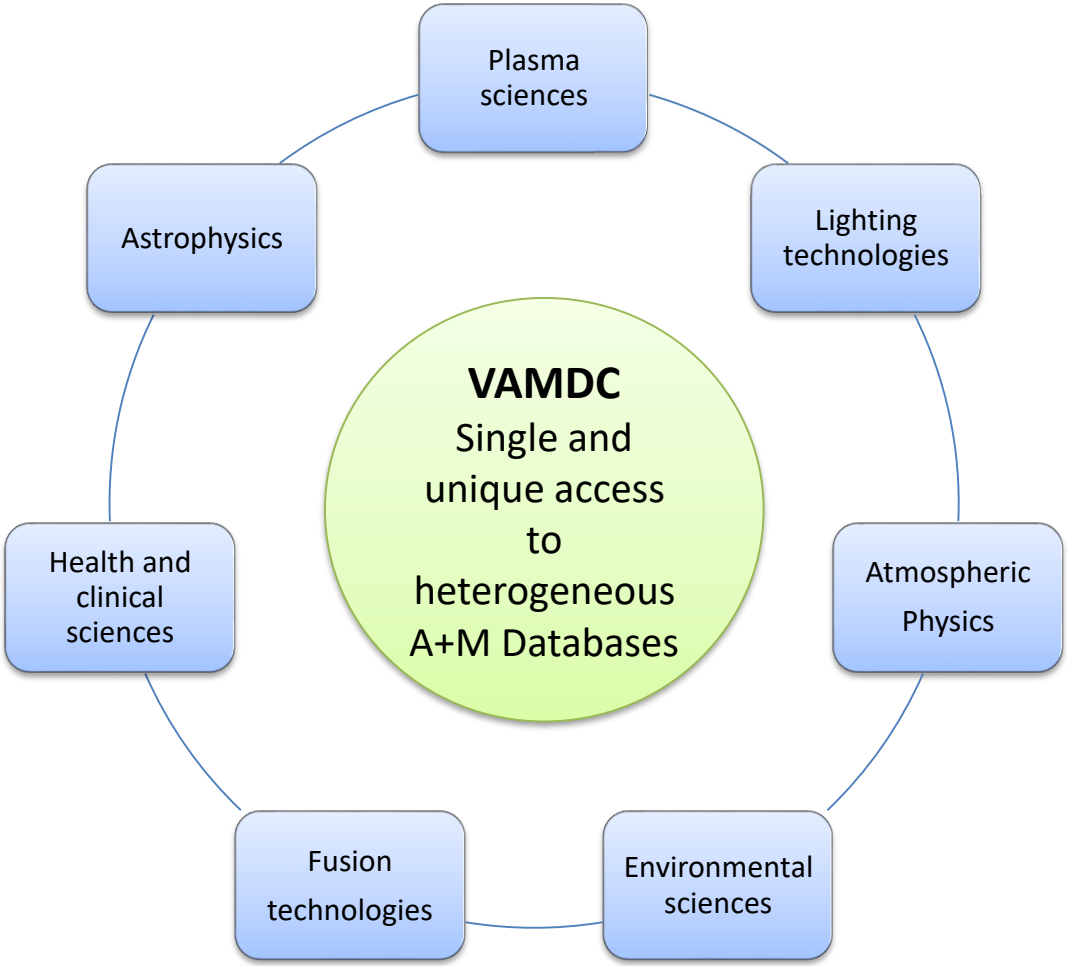
From RDA Data Citation Recommendations to new paradigms for citing data from VAMDC

C.M. Zwölf and VAMDC consortium

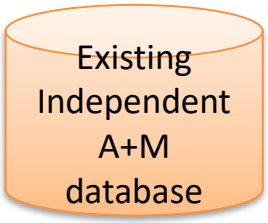
Trieste Interop – October 2016

The Virtual Atomic and Molecular Data Centre

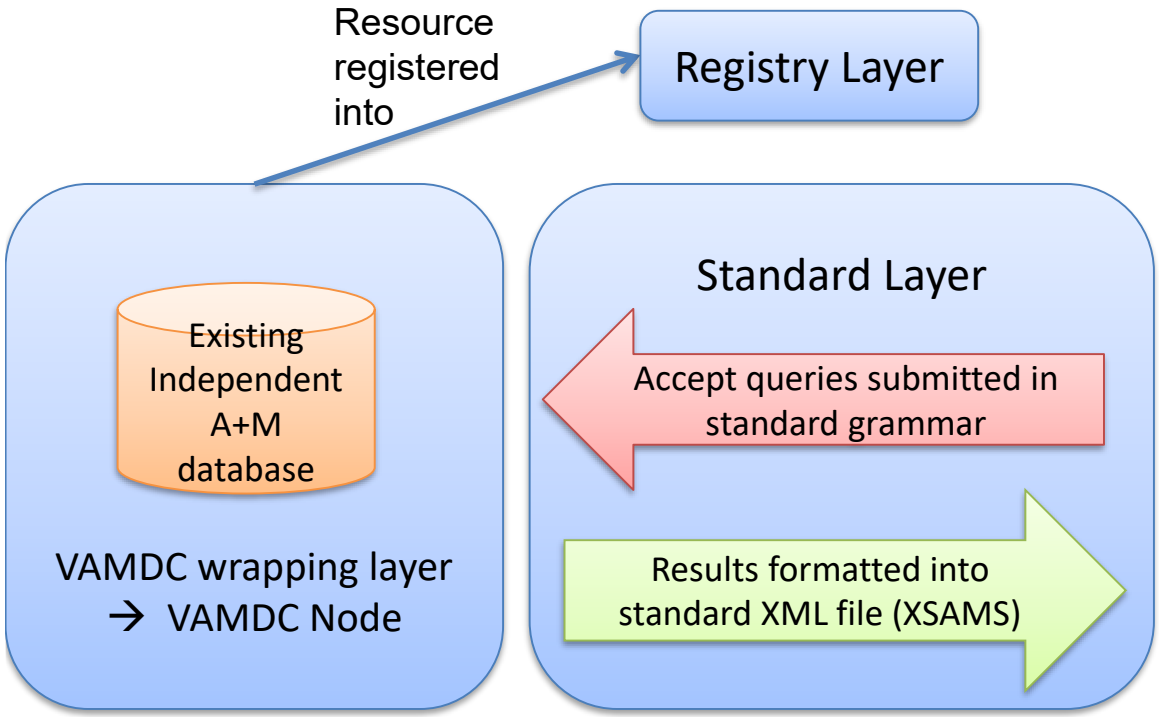


- Federates 29 heterogeneous databases http://portal.vamdc.org/
- The "V" of VAMDC stands for Virtual in the sense that the e-infrastructure does not contain data. The infrastructure is a wrapping for exposing in a unified way a set of heterogeneous databases.
- The consortium is politically organized 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

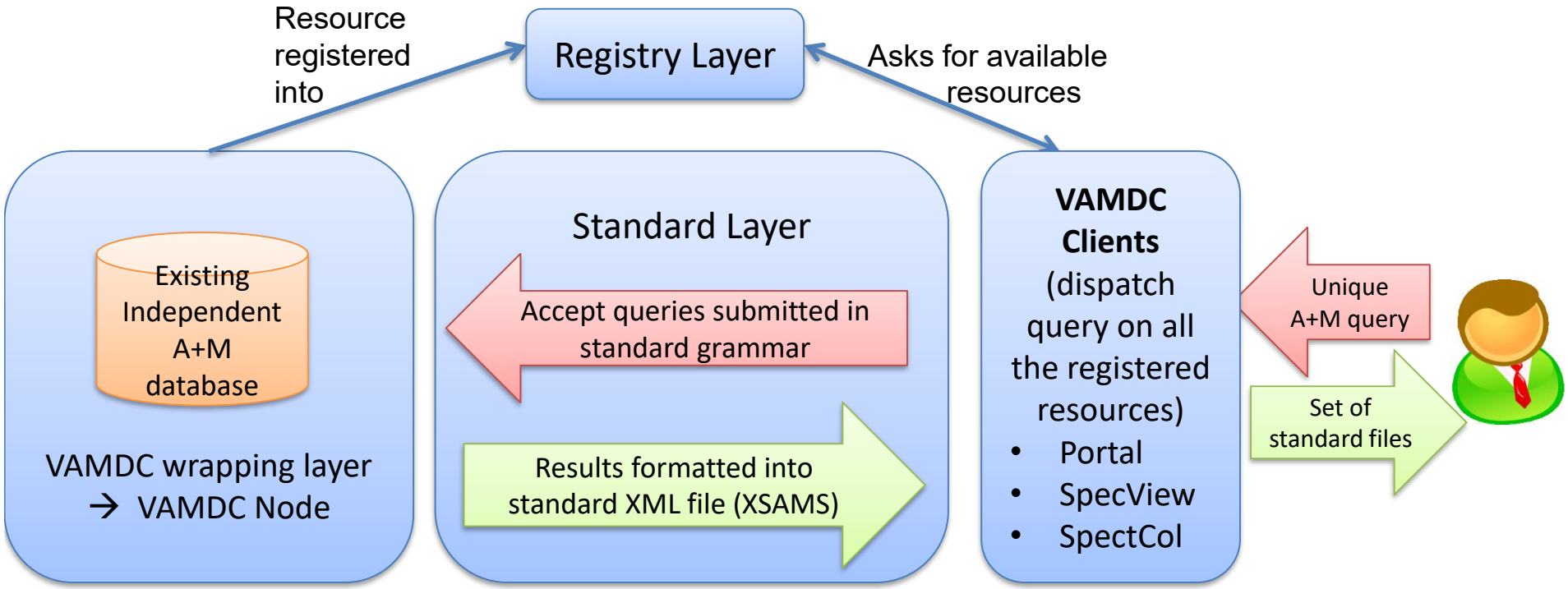
The VAMDC infrastructure technical architecture



The VAMDC infrastructure technical architecture



The VAMDC infrastructure technical architecture



Importance of citation in building new knowledge

Citation is a key element in the production of new knowledge

Gives credits to the author of the intellectual product cited

VAMDC consortium

Importance of citation in building new knowledge

Citation is a key element in the production of new knowledge

- Gives credits to the author of the intellectual product cited
- Makes the processes described into the citing article reproducible

Importance of citation in building new knowledge

Citation is a key element in the production of new knowledge

- Gives credits to the author of the intellectual product cited
- Makes the processes described into the citing article reproducible
- Enhance trust: the new results are based on proven/ solid bases. Each author does not need to prove again an used result

Importance of citation in building new knowledge

Citation is a key element in the production of new knowledge

- Gives credits to the author of the intellectual product cited
- Makes the processes described into the citing article reproducible
- Enhance trust: the new results are based on proven/ solid bases. Each author does not need to prove again an used result

The nowadays adopted citation model works well for papers.
It cannot be easily transposed to citation of digital data…

VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

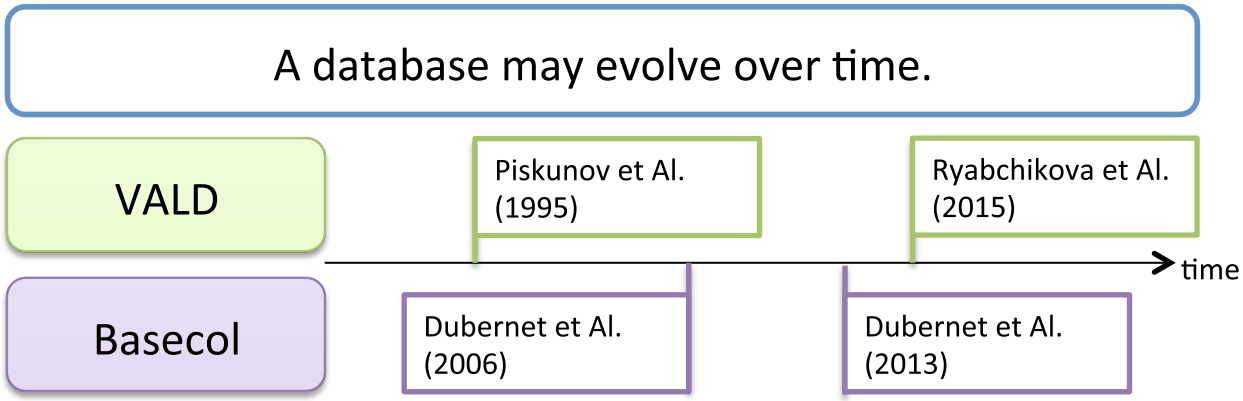
- Is very rapid
- Is not systematically reported

VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

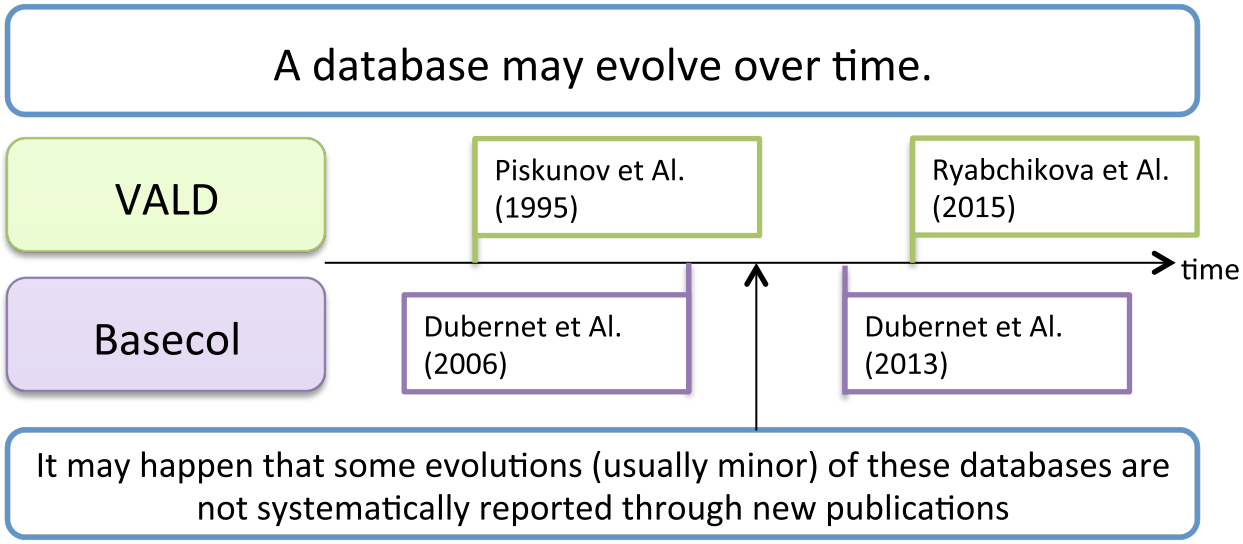


VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported



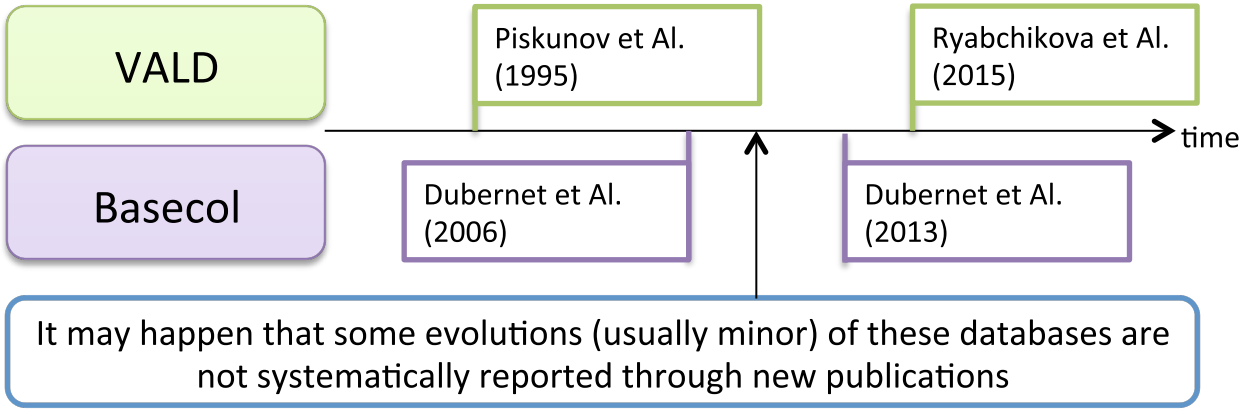
VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

A database may evolve over time.



It may happen that some evolutions (usually minor) of these databases are not systematically reported through new publications

A huge number of digital data are used nowadays in papers.

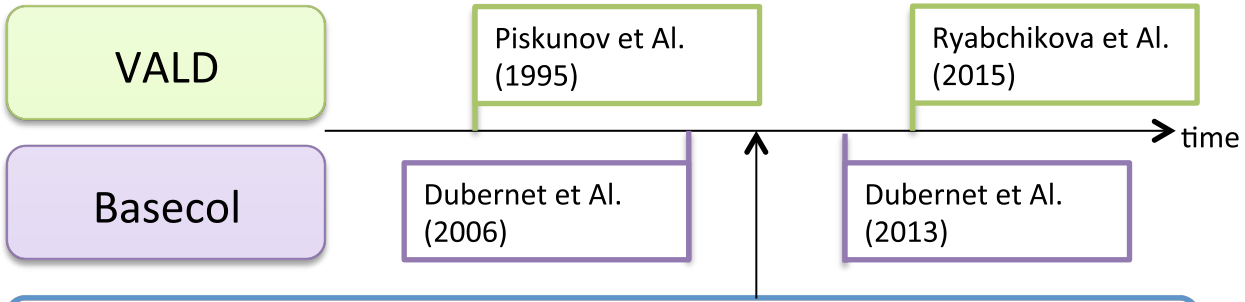
VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

A database may evolve over time.



It may happen that some evolutions (usually minor) of these databases are not systematically reported through new publications

A huge number of digital data are used nowadays in papers.

The volume of digital data is wide and constantly growing.

A given surveys may use thousands of spectroscopic data coming from many experimental/theoretical authors.

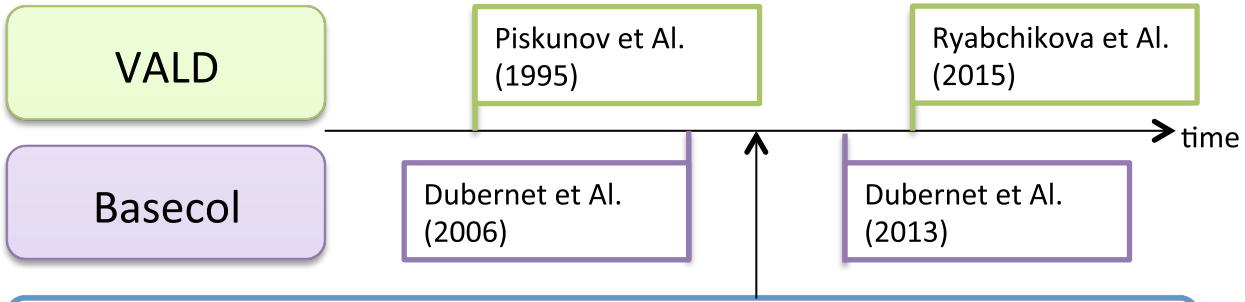
VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

A database may evolve over time.



It may happen that some evolutions (usually minor) of these databases are not systematically reported through new publications

A huge number of digital data are used nowadays in papers.

The volume of digital data is wide and constantly growing.

A given surveys may use thousands of spectroscopic data coming from many experimental/theoretical authors.

It is impossible to effectively cite the origin of thousand of data with the required fine grained granularity.

VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

A huge number of digital data are used nowadays in papers.

Citation of data is incompatible with the hand-made classic citation mechanisms.

VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

A huge number of digital data are used nowadays in papers.

Citation of data is incompatible with the hand-made classic citation mechanisms.

The survey by *[Ginard et al. (2012)]* covers frequencies from 83302Mhz to 262404Mhz detecting emission from about 36 species:

- They used catalogues from two public databases *[Picket et al. (1998)]* and *[Müller et al (2005)]* and a private communication from J. Cernicharo.
- There is no knowledge of the exact dataset used → **Their analysis is not reproducible**.
- There is no citation of the authors who produced the spectroscopic data used in their analysis.
- The collisional data are properly cited.
  - Dozen of papers for collisional data vs. hundreds of papers for spectroscopic data.

VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

A huge number of digital data are used nowadays in papers.

Citation of data is incompatible with the hand-made classic citation mechanisms.

Track the versioning of data

Having a mechanisms to speed up the citation process

VAMDC consortium

Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

A huge number of digital data are used nowadays in papers.

Citation of data is incompatible with the hand-made classic citation mechanisms.

Track the versioning of data

Having a mechanisms to speed up the citation process

- Address these issues at the VAMDC federated level (not database by database)
- Discuss these issues at the data-community level: we joined (spring 2014) the **RDA Data Citation Working Group**.

VAMDC has become one of the RDA use-cases.

VAMDC consortium

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 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 defines differently 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 defines differently 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 implement the recommendation!!

Implementation will be 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**.

Let us implement the recommendation!!

Implementation will be 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

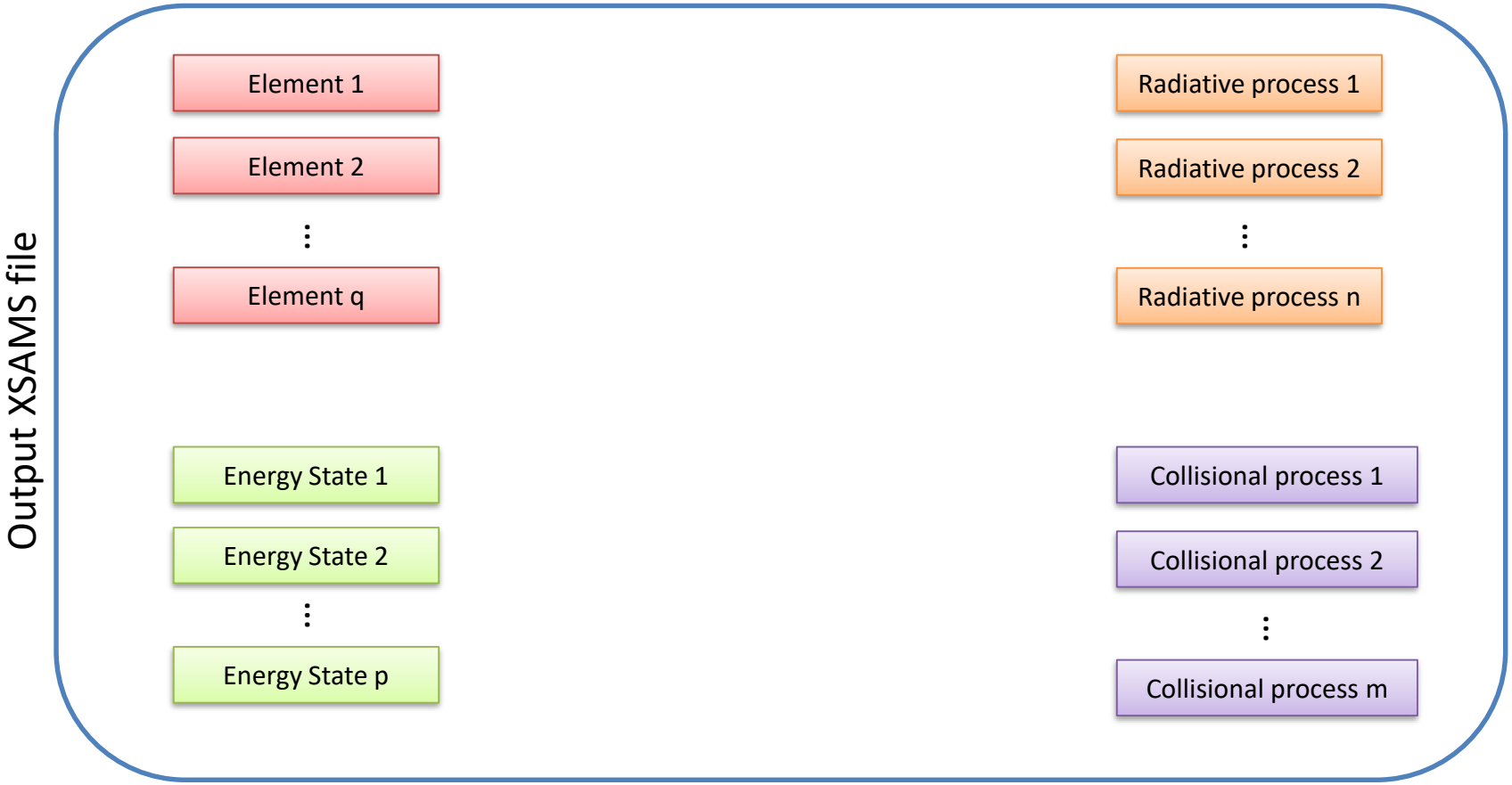
Is built over the versioning of Data

Is plugged over the existing VAMDC data-extraction mechanisms.

Due to the distributed VAMDC architecture, the Query Store architecture is similar to a log-service.

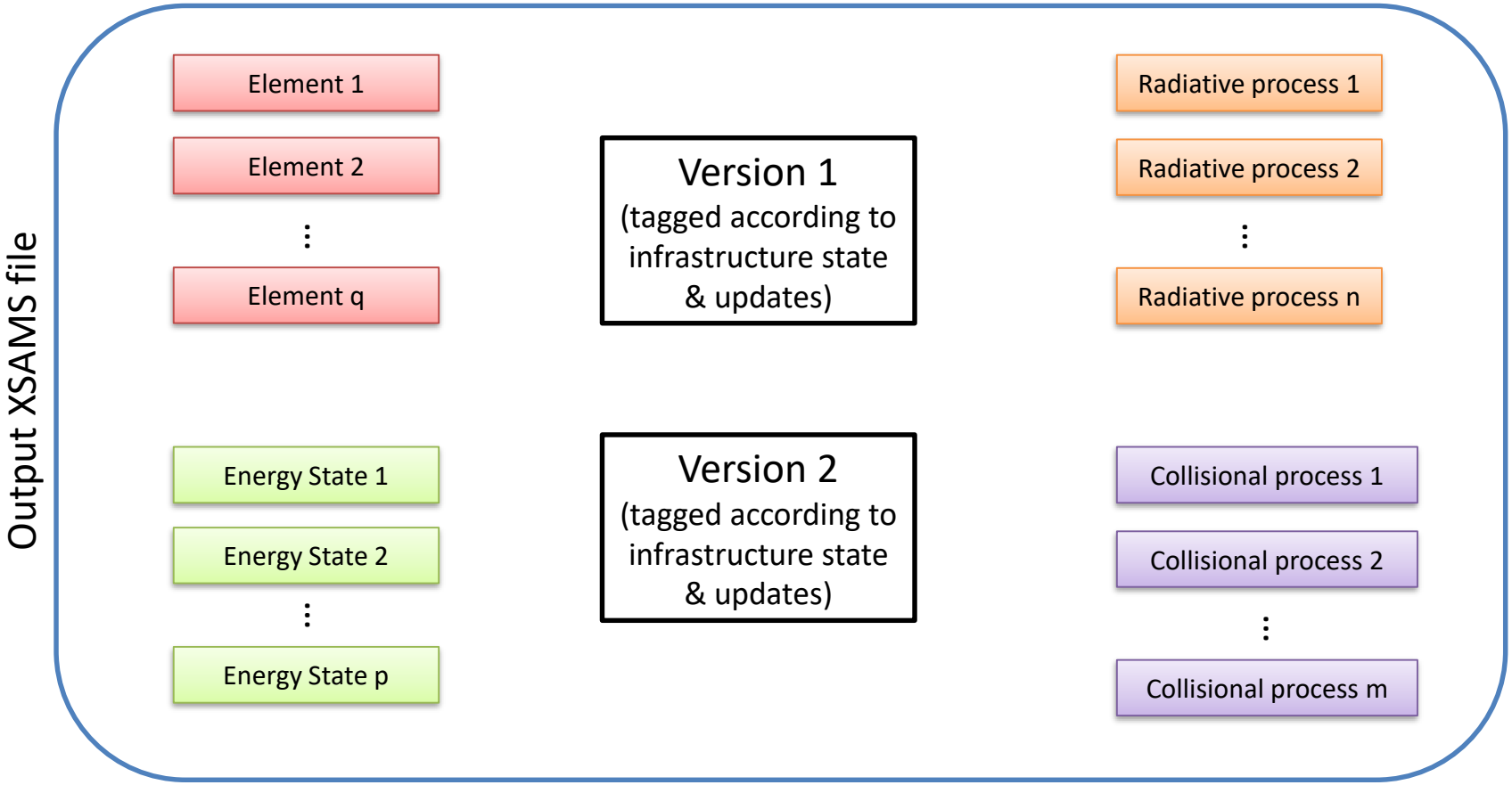
Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigm (weak structuration):



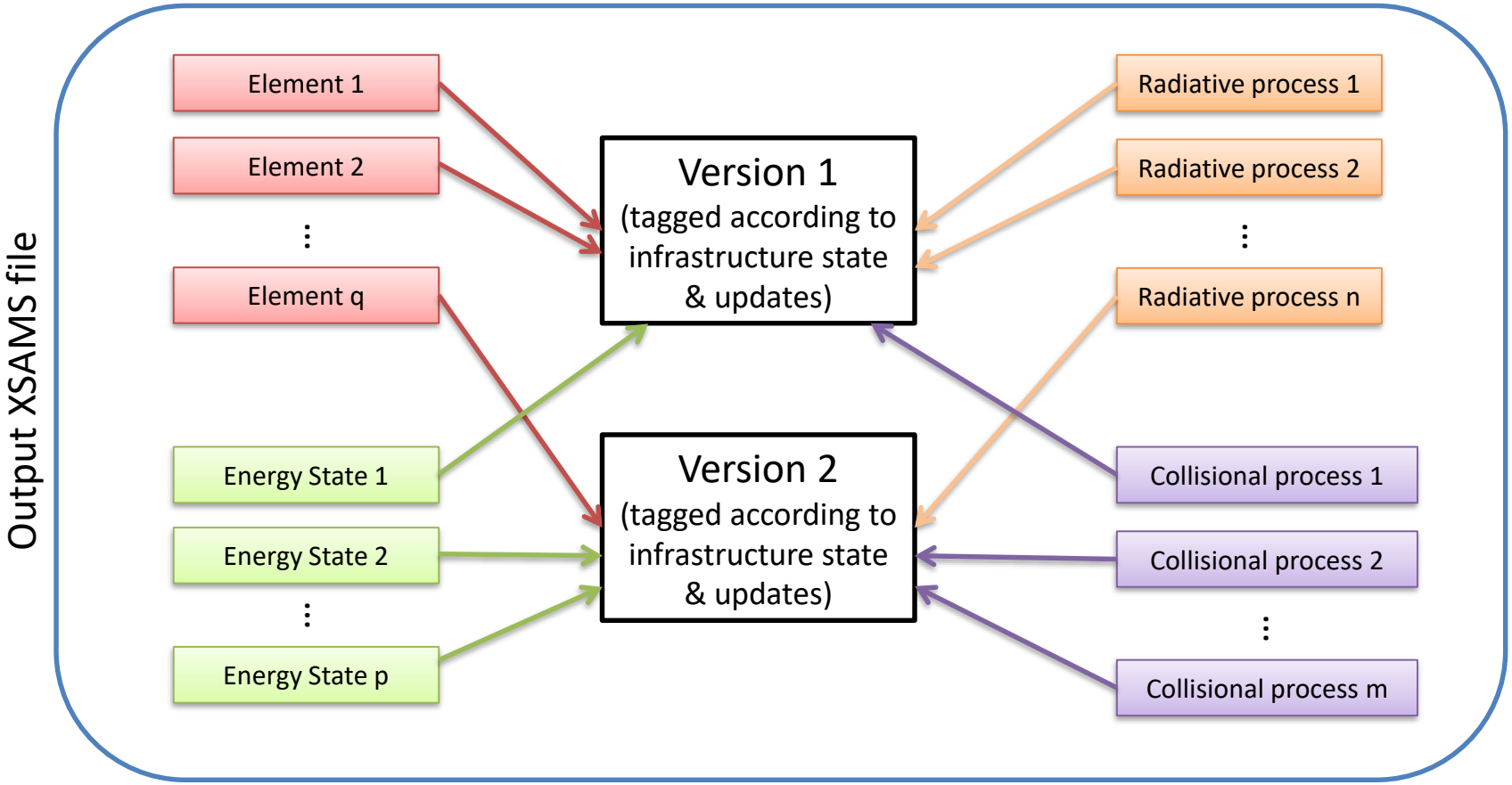
Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigm (weak structuration):



Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigm (weak structuration):



Data-Versioning: overview of the fine grained mechanisms

We adopted a change of paradigm (weak structuration):

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)

Data-Versioning: overview of the fine grained mechanisms

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 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

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.

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 joint collaboration between VAMDC and RDA-Europe.

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

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.

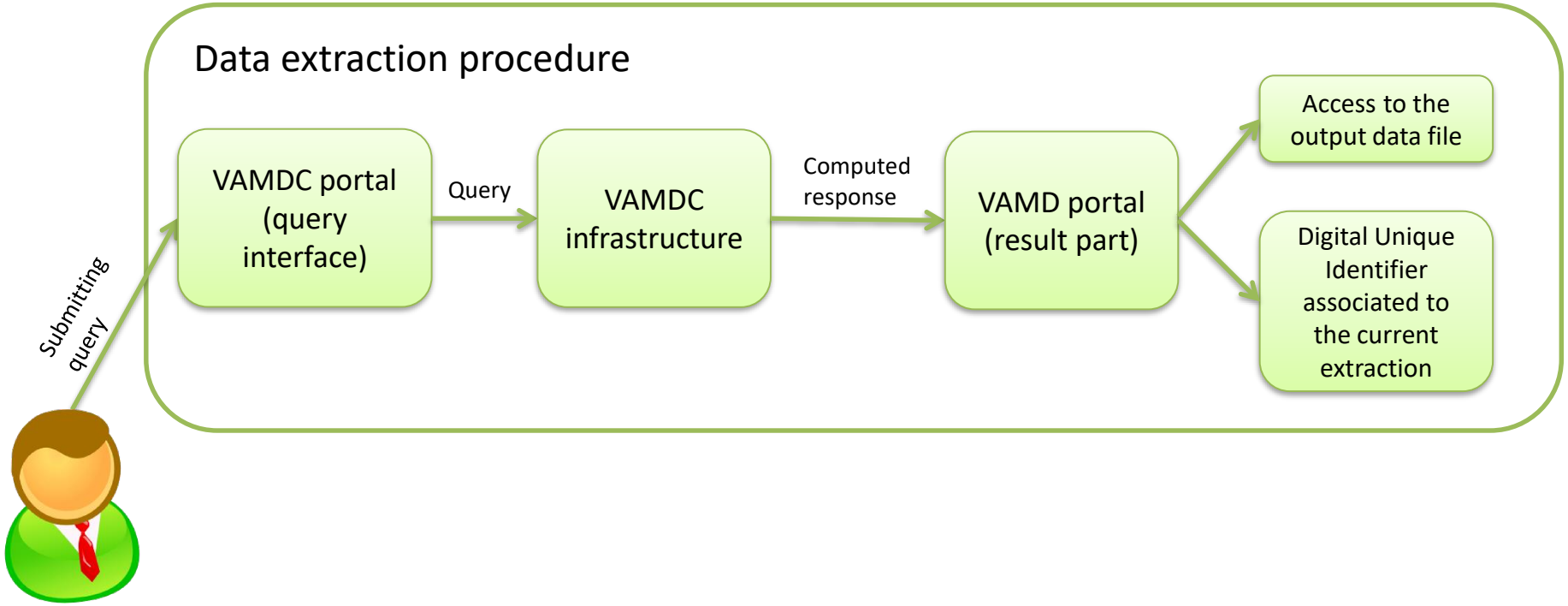
- Development started during spring 2016.
- Final product released during 2017.

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)

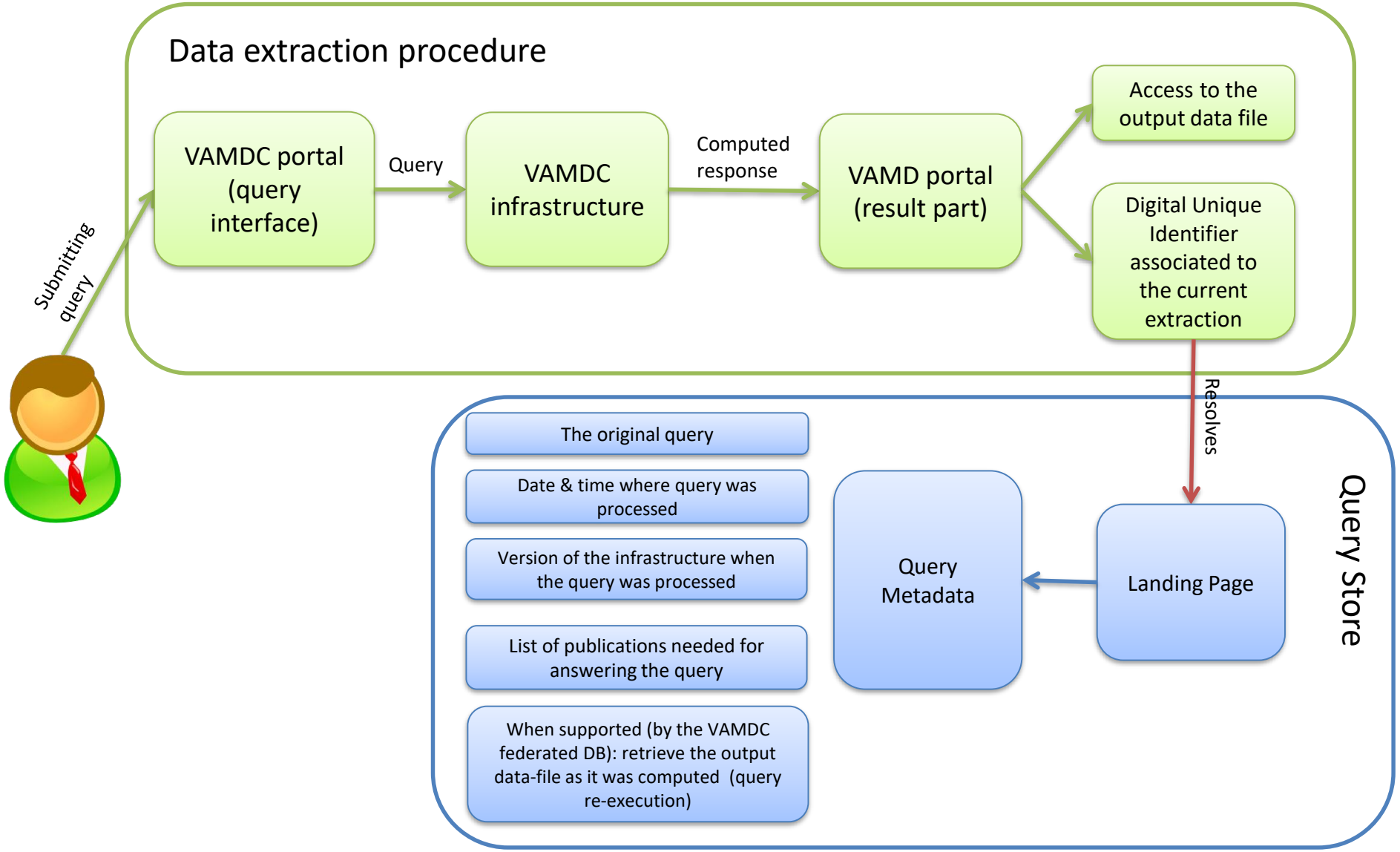
Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



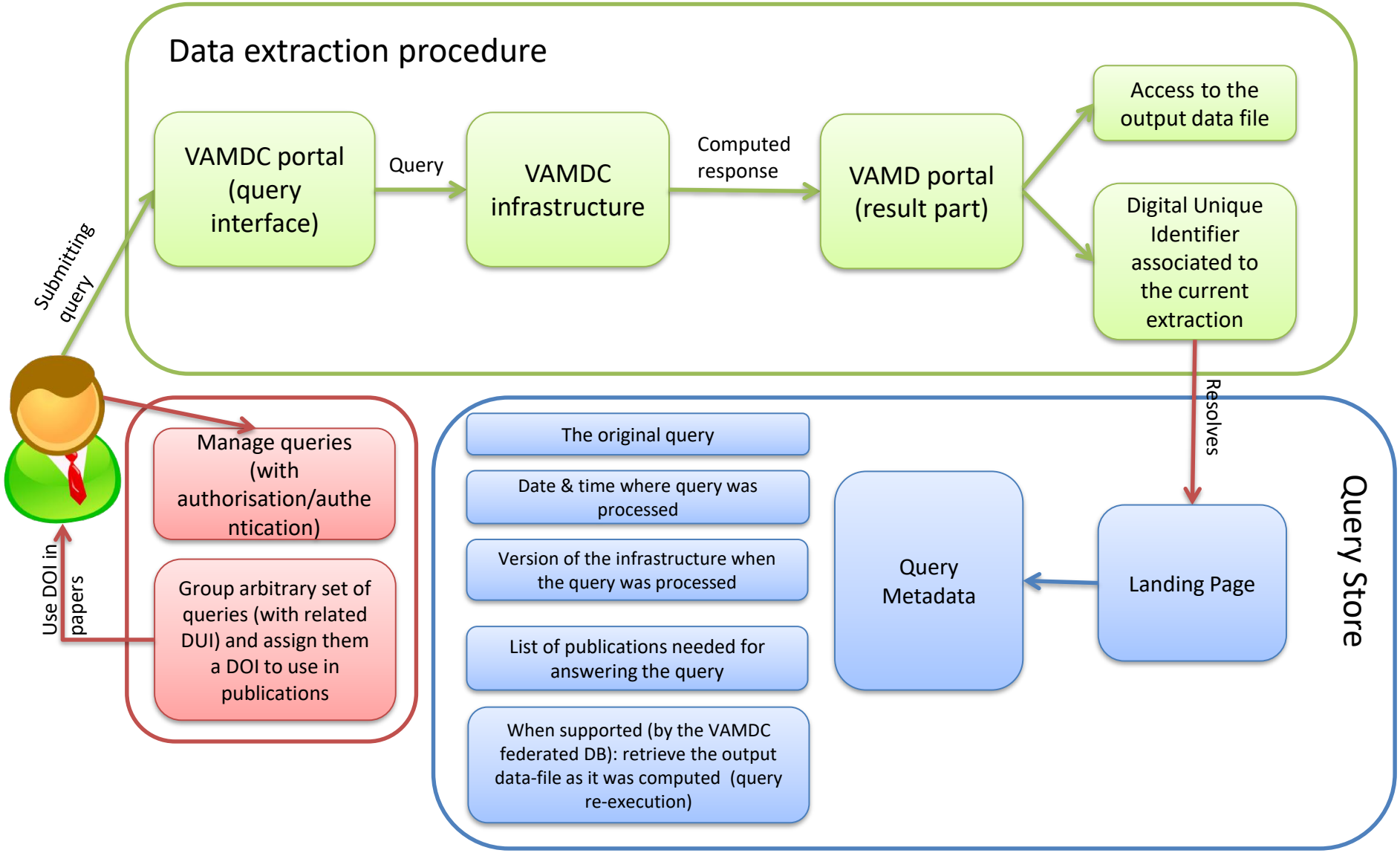
Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



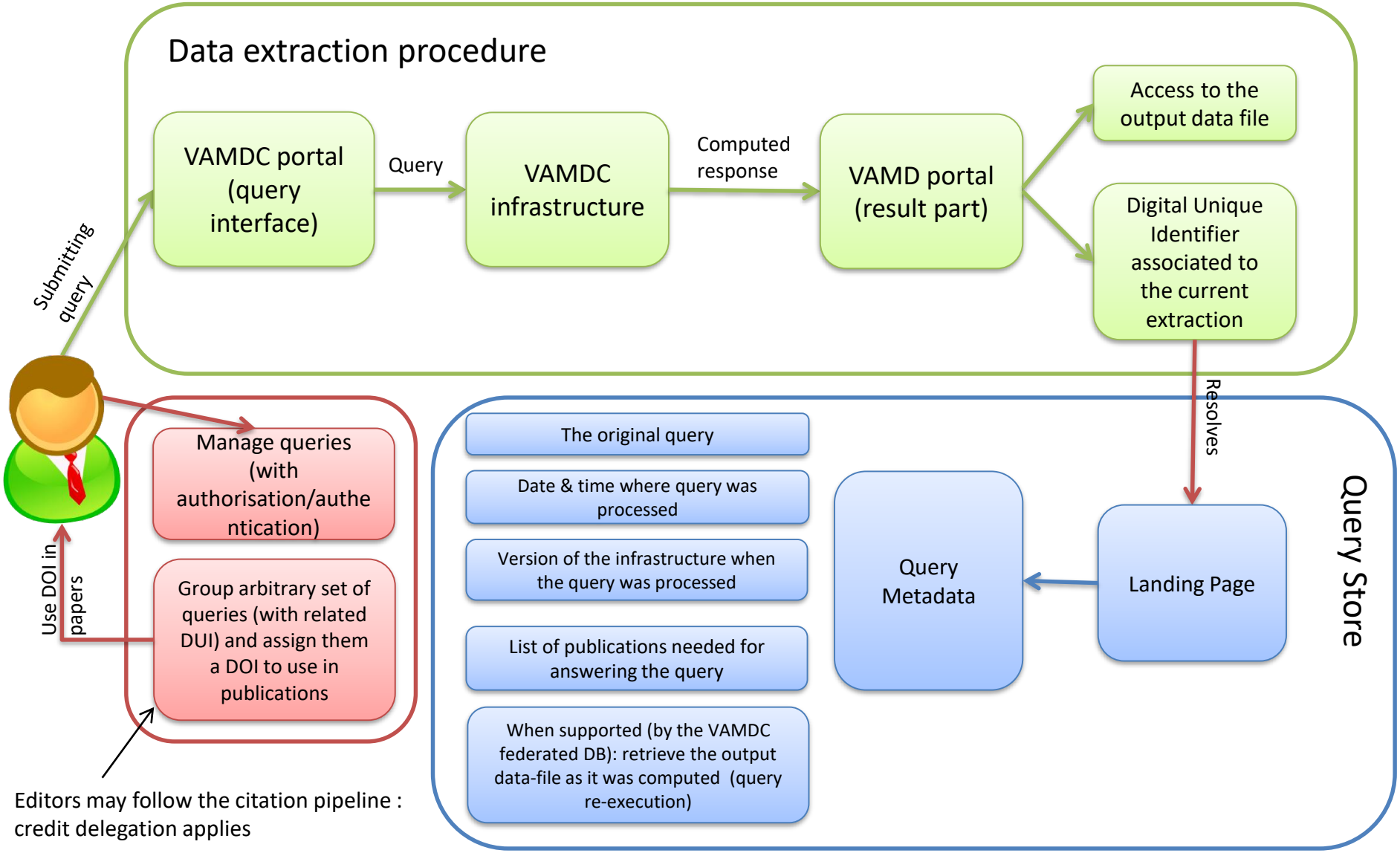
Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



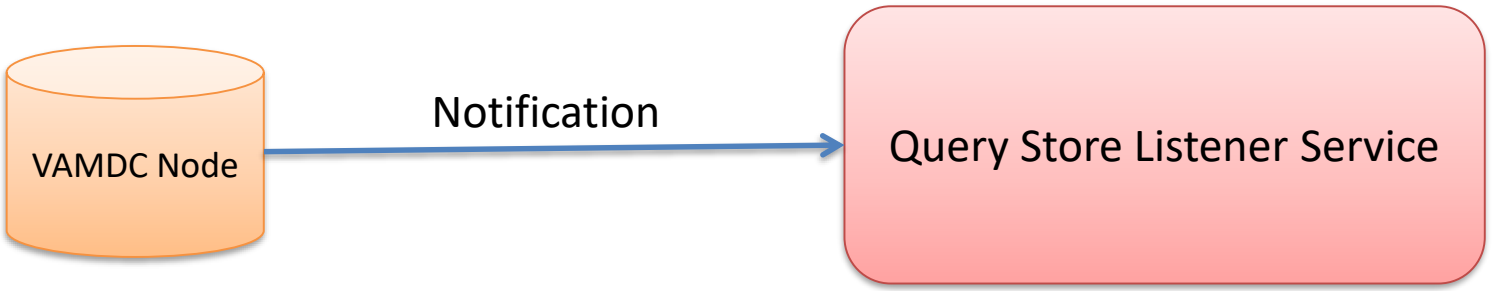
Let us focus on the query store:

Sketching the functioning – From the final-user point of view:



Let us focus on the query store:

Sketching the functioning – Technical internal point of view:

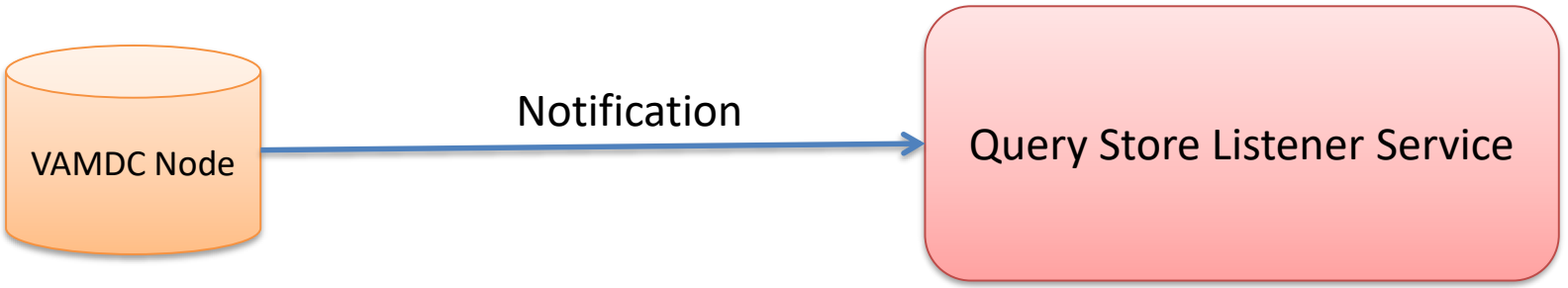


1 → When a node receives a user query, it notifies to the Listener Service the following information:

- The identity of the user (optional)
- The used client software
- The identifier of the node receiving the query
- The version (with related timestamp) of the node receiving the query
- The version of the output standard used by the node for replying the results
- The query submitted by the user
- The link to the result data.

Let us focus on the query store:

Sketching the functioning – Technical internal point of view:

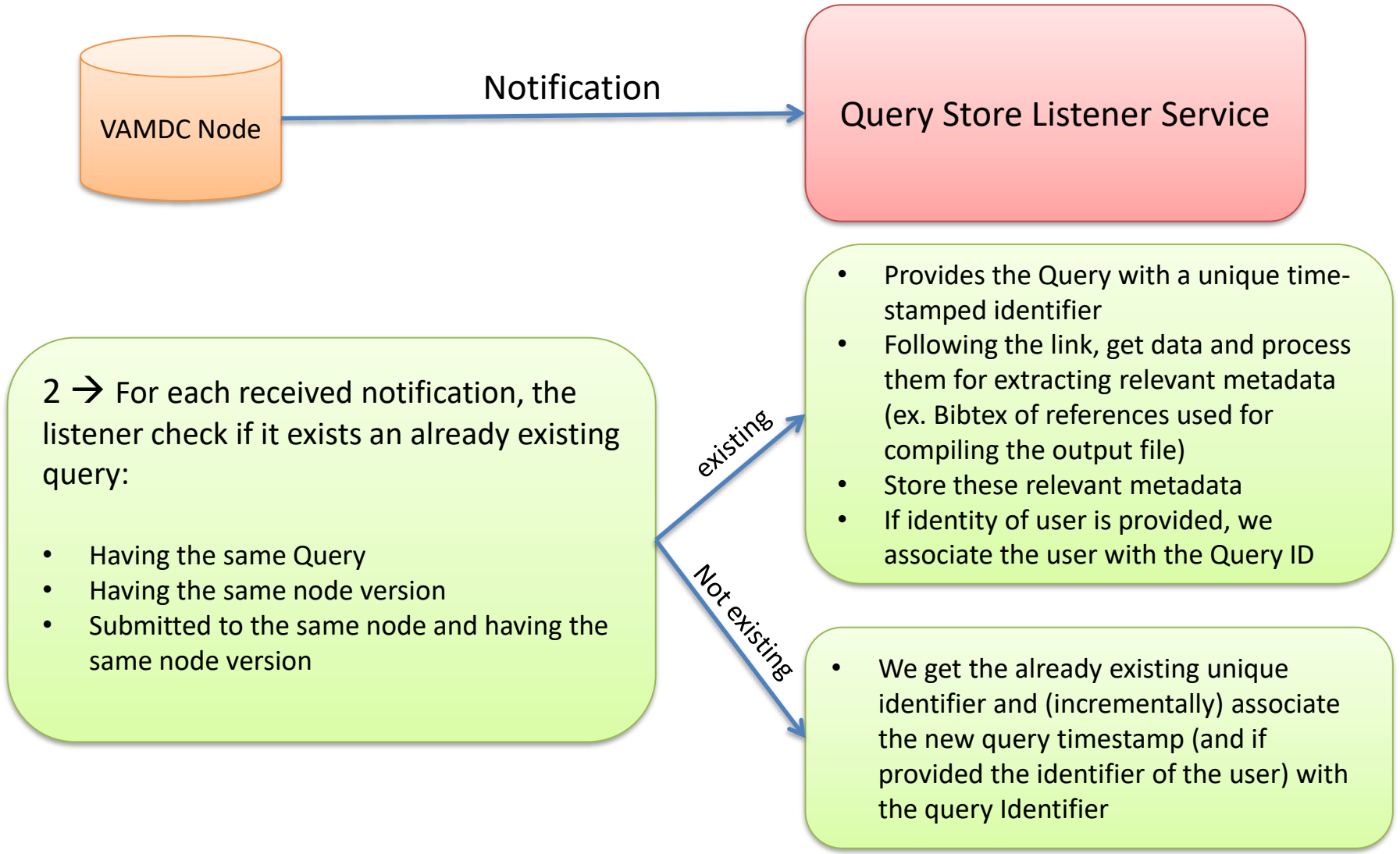


2 → For each received notification, the listener check if it exists an already existing query:

- Having the same Query
- Having the same node version
- Submitted to the same node and having the same node version

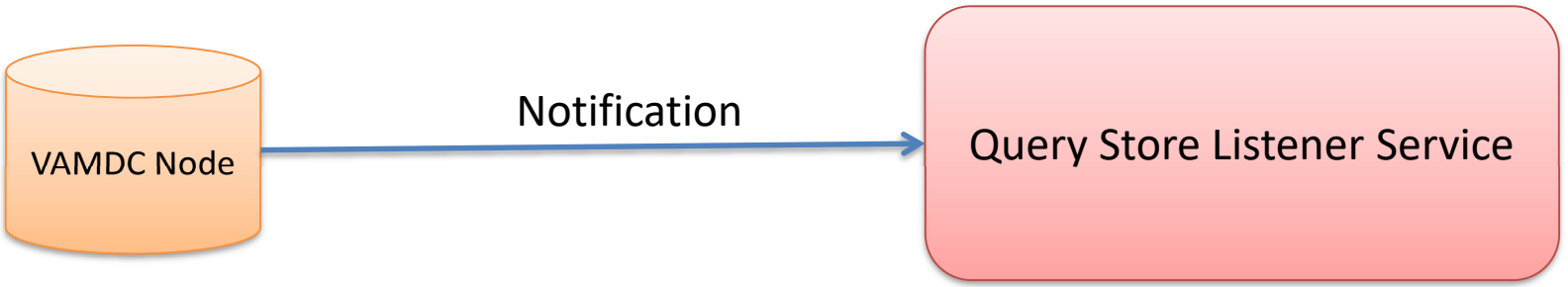
Let us focus on the query store:

Sketching the functioning – Technical internal point of view:



Let us focus on the query store:

Sketching the functioning – Technical internal point of view:



Remark on query uniqueness:

- The query language supported by the VAMDC infrastructure is VSS2 (VAMDC SQL Subset 2, http://vamdc.eu/documents/standards/queryLanguage/vss2.html).
- We are working on a specific VSS2 parser (based on Antlr) which should identify, from queries expressed in different ways, the ones that are semantically identical
- We are designing this analyzer as an independent module, hoping to extend it to all SQL.

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)