

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

C.M. Zwölf and VAMDC consortium
Heidelberg June 2016



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

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...

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

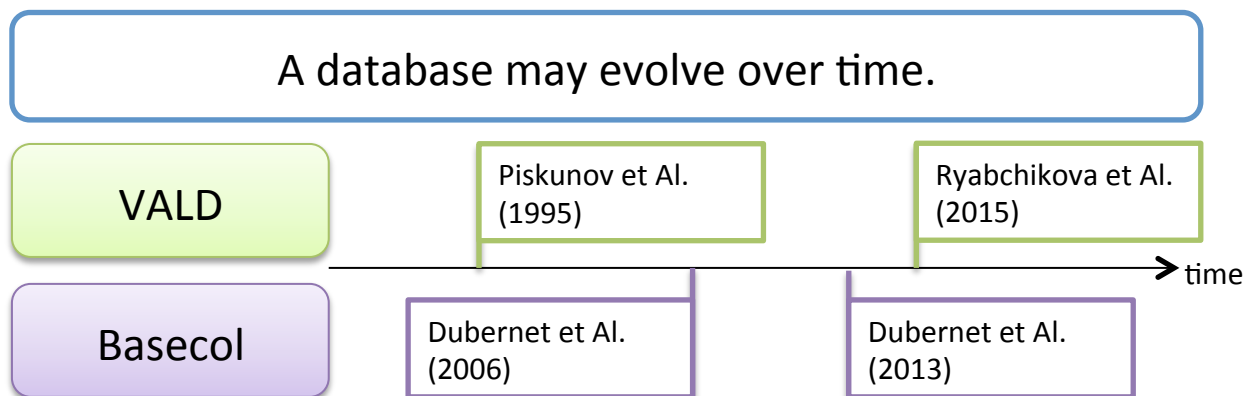
The evolution of digital data:

- Is very rapid
- Is not systematically reported

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

The evolution of digital data:

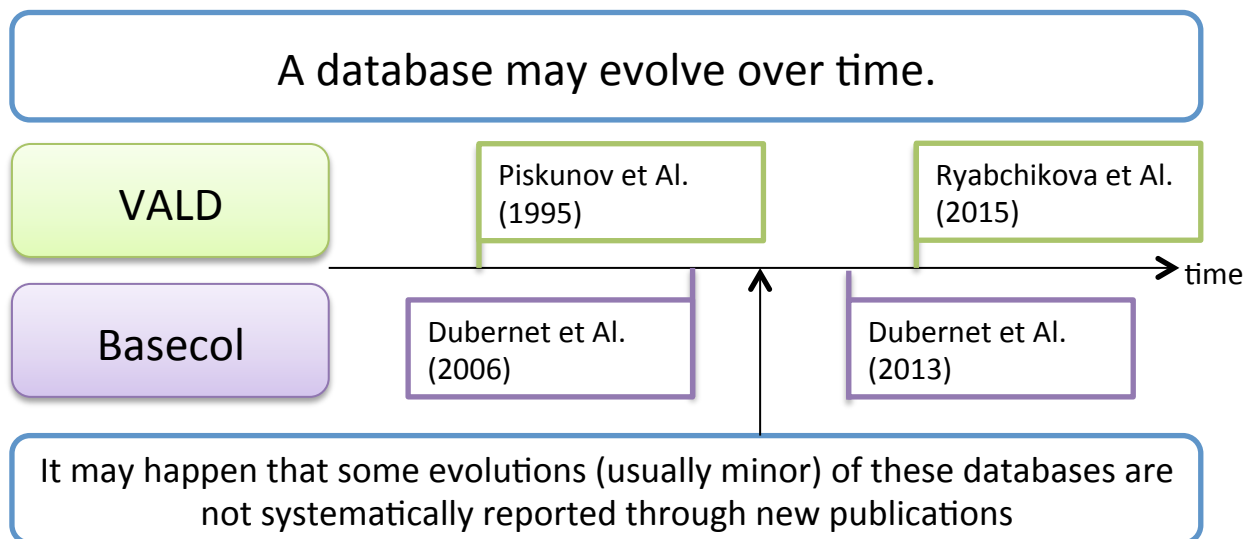
- Is very rapid
- Is not systematically reported



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

The evolution of digital data:

- Is very rapid
- Is not systematically reported

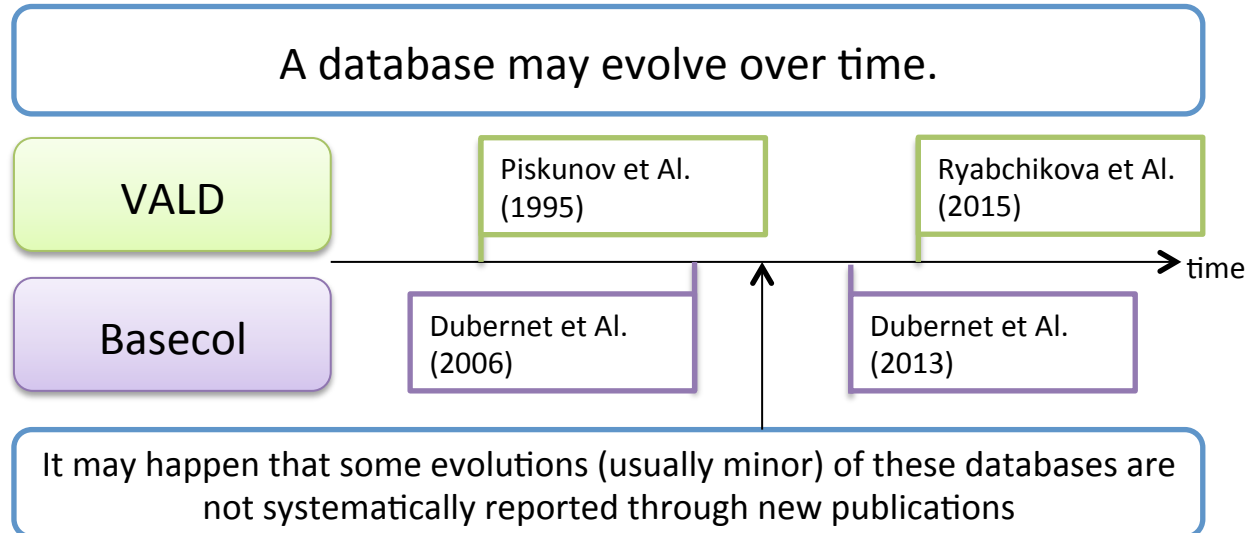


Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

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



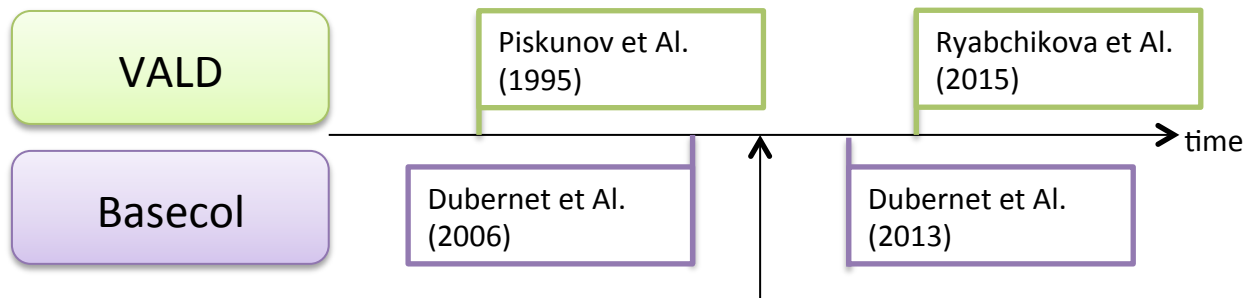
Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

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

A database may evolve over time.



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

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.

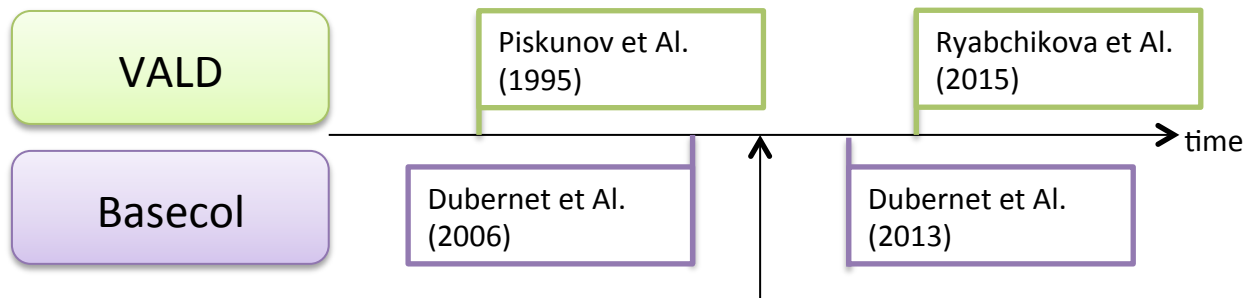
Issues in data citation: case of the Atomic and Molecular data

The evolution of digital data:

- Is very rapid
- Is not systematically reported

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

A database may evolve over time.



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

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.

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.

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.

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

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.

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 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 come from standalone databases or warehouses.
- 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!

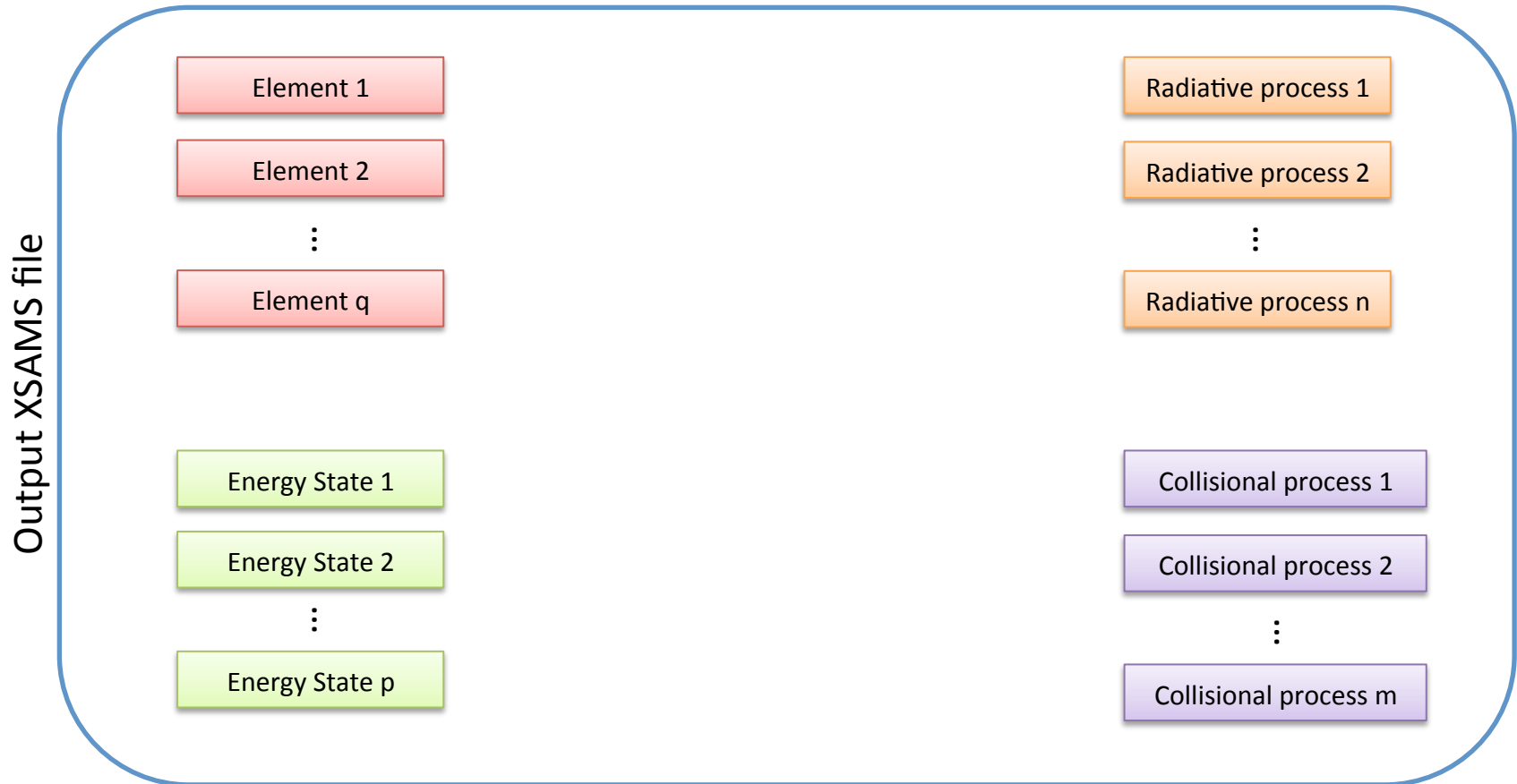
Yes, but everyone has its own definition

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.

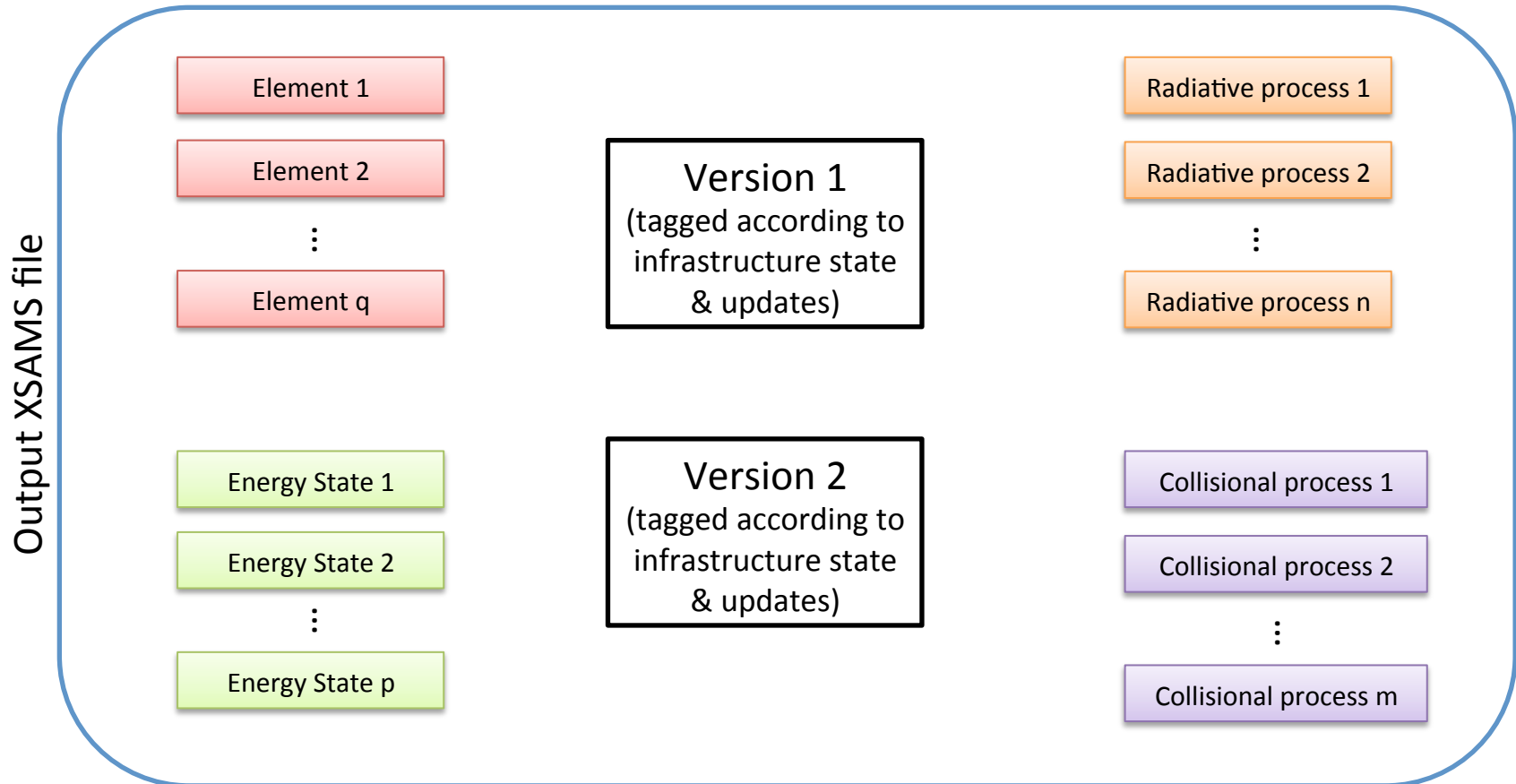
Let us focus on data tagging/versioning issue:

We adopted a change of paradigms (weak structuration):



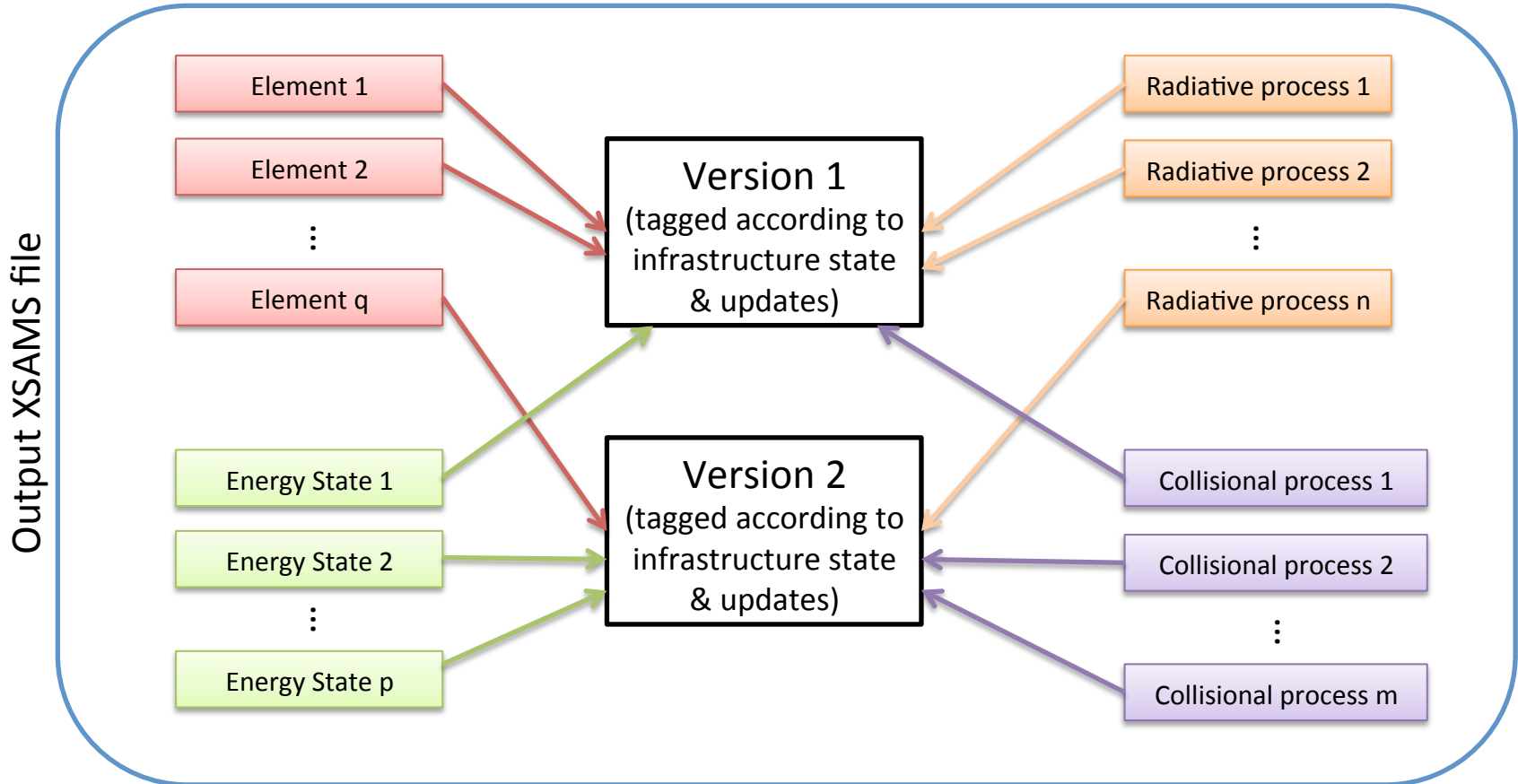
Let us focus on data tagging/versioning issue:

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 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 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:

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.

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

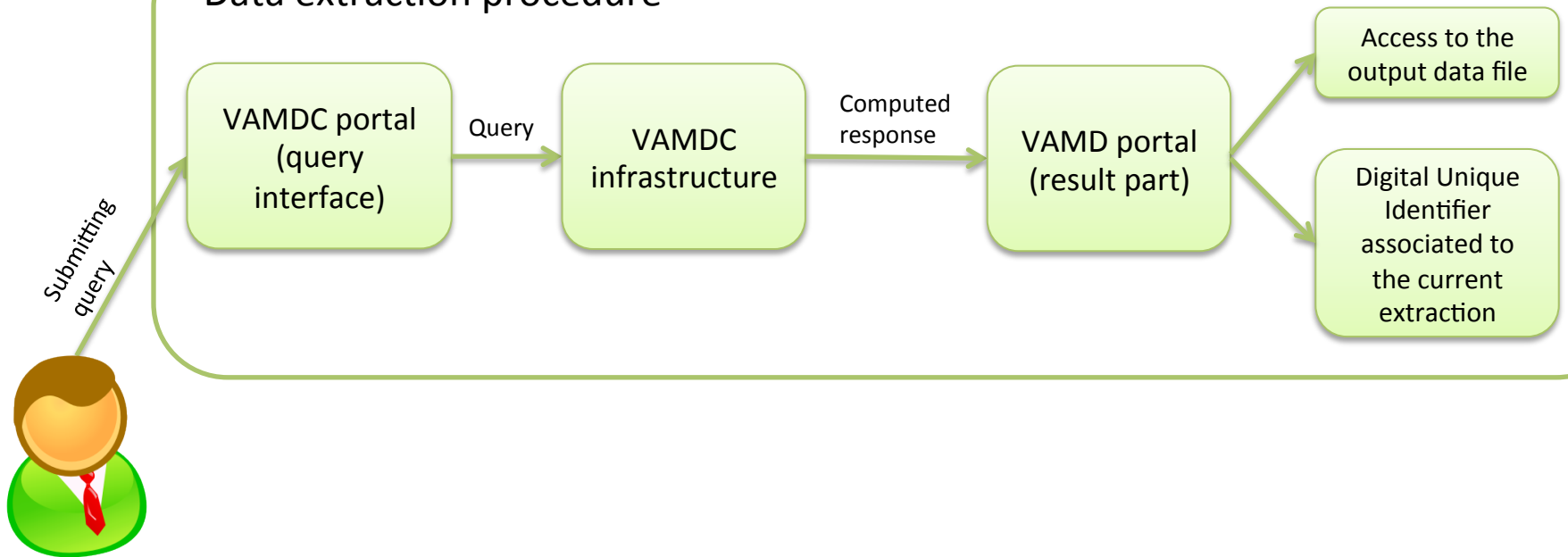
Designing technical solution for

- 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:

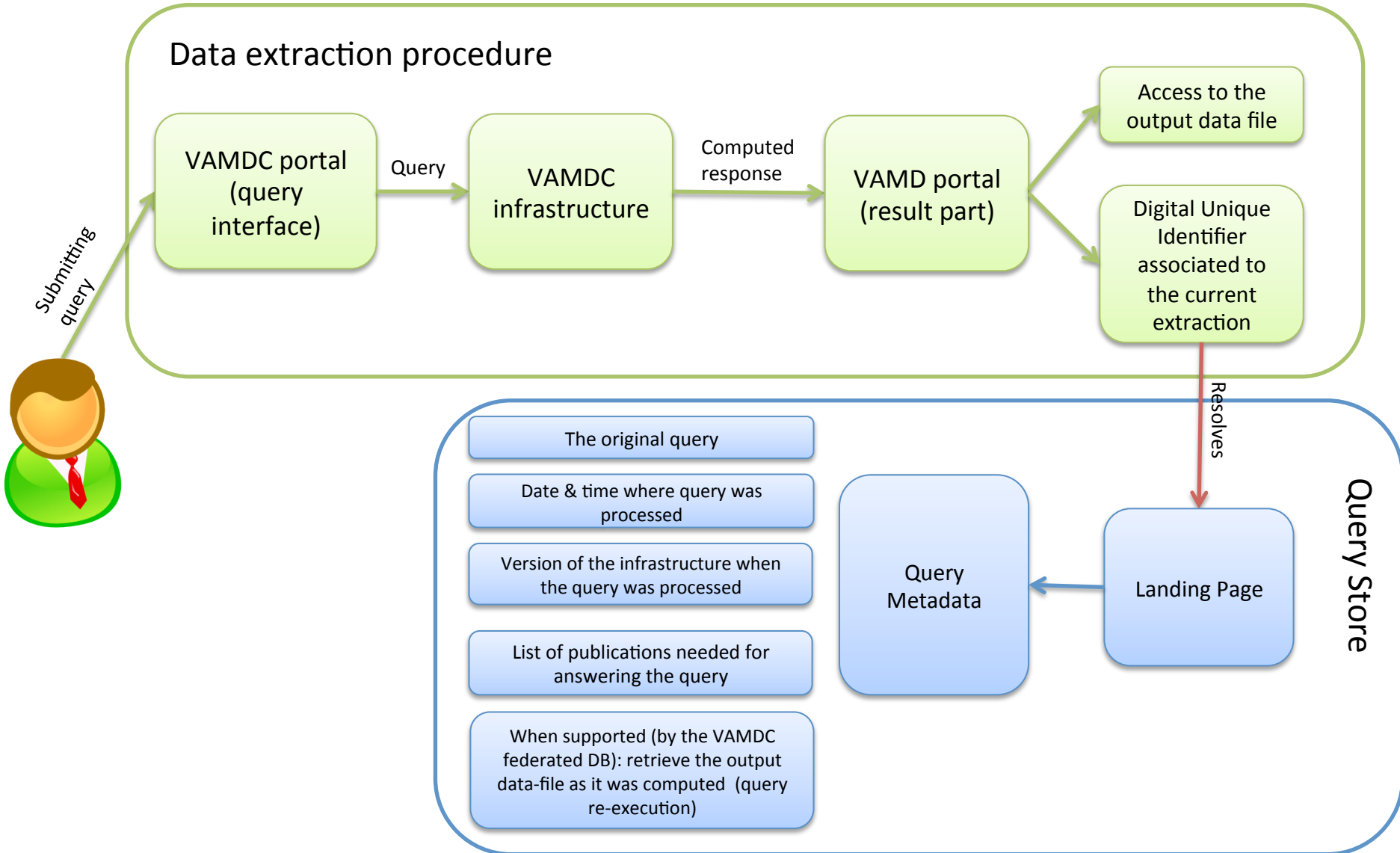
Sketching the functioning:

Data extraction procedure



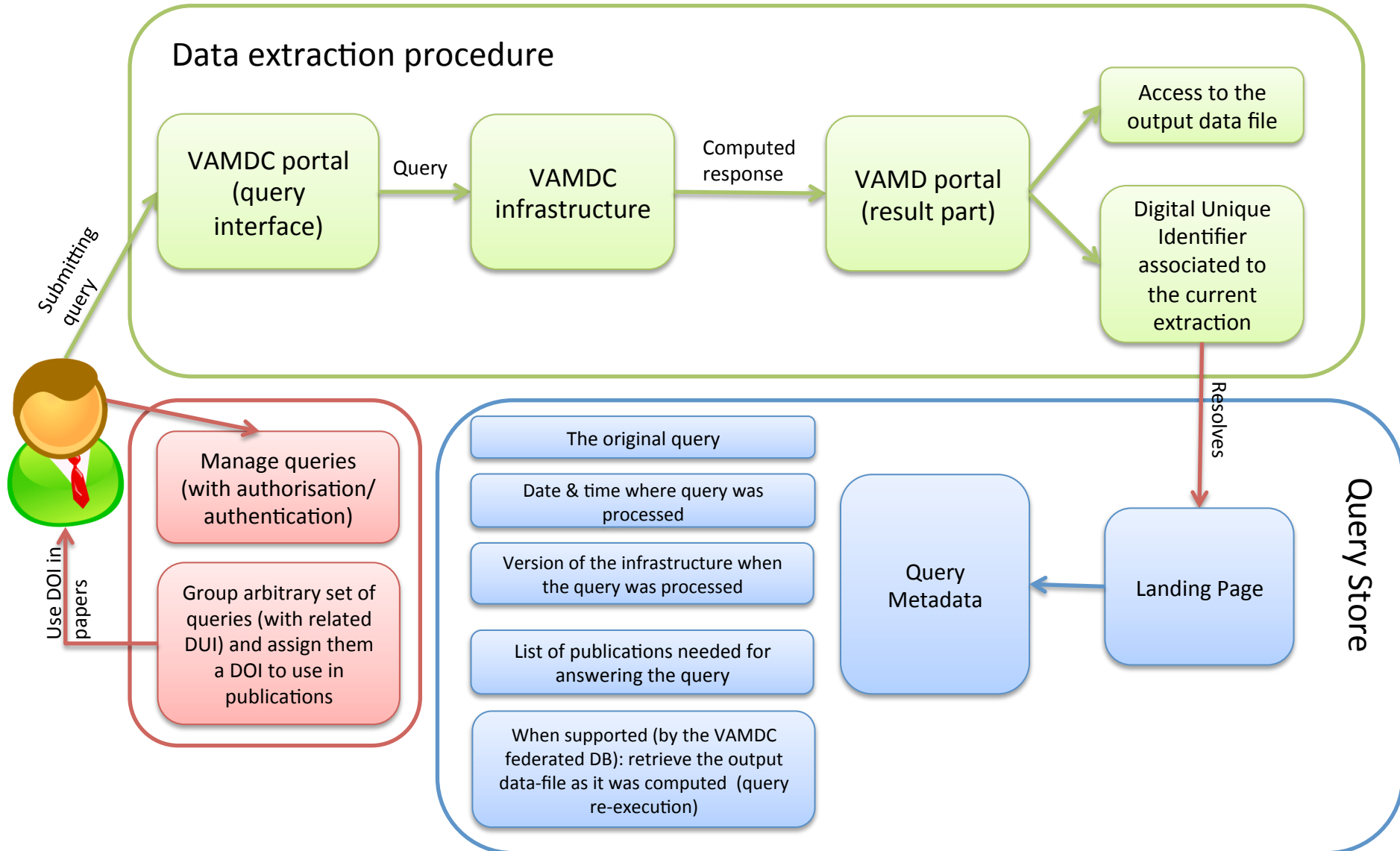
Let us focus on the query store:

Sketching the functioning:



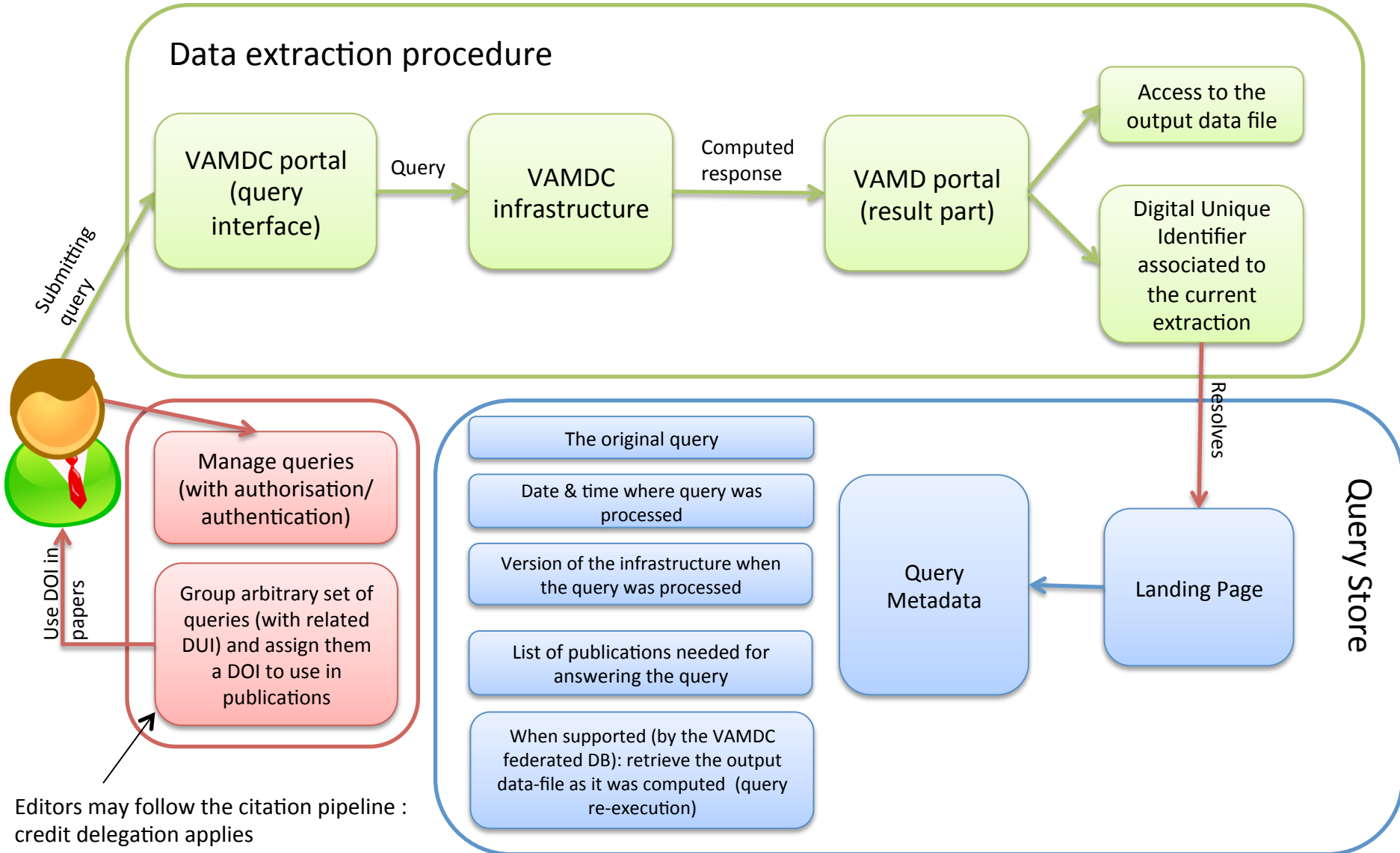
Let us focus on the query store:

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)

Bibliography (in order of citation)

Dubernet, M.L., Grosjean, A., Flower, D., Roueff, E., Daniel, F., Moreau, N., Debray, B., 2006. Ro-Vibrational Collisional Excitation Database: BASECOL - <http://www.obspm.fr/basecol>. Journal of Plasma Research SERIES 7, 356–357.

Dubernet, M.L., Alexander, M.H., Ba, Y.A., Balakrishnan, N., Balança, C., Ceccarelli, C., Cernicharo, J., Daniel, F., Dayou, F., Doronin, M., Dumouchel, F., Faure, A., Feautrier, N., Flower, D.R., Grosjean, A., Halvick, P., Klos, J., Lique, F., McBane, G.C., Marinakis, S., Moreau, N., Moszynski, R., Neufeld, D.A., Roueff, E., Schilke, P., Spielfeld, A., Stancil, P.C., Stoecklin, T., Tennyson, J., Yang, B., Vasserot, A.M., Wiesenfeld, L., 2013. BASECOL2012: A collisional database repository and web service within the Virtual Atomic and Molecular Data Centre (VAMDC). *Astron. & Astrophys.* 553, A50. doi:10.1051/0004-6361/201220630.

Ryabchikova, T., Piskunov, N., Kurucz, R.L., Stempels, H.C., Heiter, U., Pakhomov, Y., Barklem, P.S., 2015. A major upgrade of the VALD database. *Physica Scripta* 90, 054005. doi:10.1088/0031-8949/90/5/054005.

Piskunov, N.E., Kupka, F., Ryabchikova, T.A., Weiss, W.W., Jeffery, C.S., 1995. VALD: The Vienna Atomic Line Data Base. *Astron. Astrophys. Suppl. Ser.* 112, 525.

Ginard, D., Gonzalez-Garcia, M., Fuente, A., Cernicharo, J., Alonso-Albi, T., Pilleri, P., Gerin, M., Garcia-Burillo, S., Ossenkopf, V., Rizzo, J. R., Kramer, C., Goicoechea, J. R., Pety, J., Bern, O., Joblin, C., 2012. Spectral line survey of the ultracompact hii region monoceros r2? *Astron. & Astrophys.* 543, A27. URL: <http://dx.doi.org/10.1051/0004-6361/201118347>, doi:10.1051/0004-6361/201118347.