PRACTICE PAPER

Implementing in the VAMDC the New Paradigms for Data Citation from the Research Data Alliance

Carlo Maria Zwölf[1,2,3,4], Nicolas Moreau[1,2,3,4], Yaye-Awa Ba[1,2,3,4] and Marie-Lise Dubernet[1,2,3,4]

[1] LERMA, Observatoire de Paris, FR

[2] PSL Research University, FR

[3] CNRS, Sorbonne University, FR

[4] UPMC Univ Paris 06, 5 Place Janssen, 92190 Meudon, FR

Corresponding author: Carlo Maria Zwölf (carlo-maria.zwolf@obspm.fr)

VAMDC bridged the gap between atomic and molecular (A&M) producers and users by providing an interoperable e-infrastructure connecting A&M databases, as well as tools to extract and manipulate those data. The current paper highlights how the new paradigms for data citation produced by the Research Data Alliance in order to address the citation issues in the data-driven science landscape, have successfully been implemented on the VAMDC e-infrastructure.

Keywords: database; data citation; Research Data Alliance; Scholix; atomic data; molecular data

1 Introduction

For the last decades, data and software have redefined the way of carrying out science (Hey et al. (2009)). The current volumes and complexity of data that are now being collected, produced and processed, and their inevitable increase require new tools, techniques and ways of working. A number of principles and best practices for the management of scientific data have arisen, and a consensus is being reached around themes such as data identification (Wittenburg et al. (2017)) or FAIR principles (Wilkinson et al. (2016)).

In this fast evolving landscape of data-intensive science, the *citation* is an anchor: it remains a key element in the production of new knowledge, since it enhances trust (the new results are based on proven/solid bases and a scientist does not need to prove again a used result), makes the process described by the cited work reproducible and gives credits to the author of the cited intellectual product. According to the FAIR principles, most of the data should be re-used in derived works: the role of *Citation* is crucial in open-data-driven science. However, the classical citation paradigm used in scientific papers (mostly hand-made bibliographies and referring to other papers) is incompatible with the current data-deluge (Bell et al. (2009)): on one hand a huge number of digital data (with disparate origins) may be used in a given paper; on the other hand the evolution of digital data is very rapid and not systematically reported.

In the context of the Virtual Atomic and Molecular Data Centre we aimed at addressing these issues at the data-community level and in 2014 we joined the Research Data Alliance. The RDA, through its Data Citation Working Group[1] and RDA/WDS Scholarly Link Exchange (Scholix) Working Group,[2] has defined new models for citation in the digital era.

In this paper, after recalling some technical elements of the VAMDC e-infrastructure (section 2.1) and the recommendations coming from the RDA-Data Citation WG and Scholix WG (respectively section 2.2 and 2.3), we focus on how these recommendations are implemented ov0er the existing VAMDC E-infrastructure

---

[1] https://www.rd-alliance.org/groups/data-citation-wg.html.

[2] https://www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg, which is a follow up of the RDA/WDS Publishing Data Services WG (https://rd-alliance.org/groups/rdawds-publishing-data-services-wg.html).

(section 3 and 4 respectively for Data-Citation and Scholix). After describing the ongoing/further works (section 5) we conclude with discussion (section 6).

2 Technical framework
2.1 The VAMDC e-infrastructure

The Virtual Atomic and Molecular Data Centre (VAMDC, (Dubernet et al. (2016))) is a political and technical framework for operating and sustaining a worldwide digital research infrastructure, built over two European FP-7 projects ((Dubernet et al. 2010); (Zwölf et al. 2014)). The e-infrastructure federates in an interoperable way about 30 heterogeneous atomic and molecular databases. By providing data producers and compilers a large dissemination platform for their works, VAMDC succeeded in removing the bottleneck between data producers and the wide body of users of that data. The "V" of VAMDC stands for "virtual" in the sense that the e-infrastructure does not contain data: it is a wrapping for exposing in a unified way a set of heterogeneous databases. An *ad hoc* generic wrapping software, called the *node-software* (Regandell et al. (2018)) transforms an autonomous database into a VAMDC federated database, called *data-node*. Each *data-node* accepts queries submitted in a standard grammar (VAMDC SQL Subset (VAMDC Consortium (2012)), a subset of SQL as it names indicates) and, by implementing an interoperable data access protocol (Dowler et al. (2010)) developed by the IVOA,[3] provides output formatted into a standard XML file (VAMDC XML Schema for Atomic Molecular and Solid Data, VAMDC-XSAMS).[4] The data-nodes are listed into a specific registry (Benson et al. (2009)), a sort of yellow pages service for discovering the VAMDC available resources. The current VAMDC registry implementation is derived from the AstroGrid project (Walton (2004)).

A user wishing to extract data from VAMDC:

- may use a VAMDC client software: when the client forms a query, the client asks the registry about the availability and relevance of the data-nodes, and then dispatches the query to the nodes. Each node produces a standard VAMDC-XSAMS file. The client collects the returned file and displays the file's content to the user.
- may submit his/her query directly to the specific node he/she wants to hit, after having discovered it on the registry.

From the technical point of view, VAMDC may be seen as a distributed architecture, with no central management system.

2.2 The RDA recommendation on dynamic data citation

The Research Data Alliance[5] and its Data Citation Working Group[6] have provided the researchers and data centers communities with recommendations to identify and cite dynamic data (Asmi et al. (2016)). The proposed solution relies on a query centric view and the set-up of a *Query Store*. Data should be stored in a versioned time-stamped manner and accessed through queries. The Query Store stores all the identified and time-stamped queries together with the relevant metadata. It also gives access to the the data produced when a given query was executed. Within the context of the RDA recommendation the term "query" has to be understood in its wider sense: it stands for any processing mechanism used to extract data from a computer-based system.

We already discussed (Zwölf et al. (2016)) how the VAMDC standards have evolved in order to meet the part of the RDA recommendation related to the versioning and to the data-timestamping. In this paper, we focus on the technical details about the implementation of the Query Store, i.e. for storing timestamped queries submitted to the VAMDC infrastructure.

2.3 The RDA Scholix recommendation

The goal of the Scholix initiative (Burton et al. (2017)) is to establish a high-level interoperability framework for exchanging information about the links between scholarly literature and data. It is an evolving lightweight set of guidelines that aims to increase interoperability and to enable an open information ecosystem. The objective is to understand systematically what data underpins literature and what literature references data. The Data-Literature Interlinking Service from OpenAIRE (DLI Service)[7] is the first exemplar aggregation

---

[3] International Virtual Observatory Alliance.
[4] https://standards.vamdc.eu/dataModel/vamdcxsams/index.html#vamdcxsamslanguage-index.
[5] https://www.rd-alliance.org.
[6] https://www.rd-alliance.org/groups/data-citation-wg.html.
[7] https://scholexplorer.openaire.eu/index.html#/api.

and query service fed by the Scholix open information ecosystem. The Scholix framework, together with the DLI aggregation, is designed to enable other 3rd party services (domain-specific aggregations, integrations with other global services, discovery tools, impact assessments etc).

3 Implementing the Query Store for the VAMDC infrastructure

The RDA Data Citation recommendation is meant for standalone data-repositories and/or for warehouses. It was both technically and politically challenging to implement the RDA recommendation in the case of the distributed VAMDC infrastructure. The solution had to deal with a lot of constraints:

- any evolution of the infrastructure automatically impacts all the connected databases (there are about 30 connected databases nowadays).
- as a consequence, the majority of the VAMDC Consortium members must validate any technological evolution of the infrastructure.

Any adopted solution must lessen the load on the existing infrastructure members and have minimal implementing costs for each *data-node* owner. These constraints suggested to embed part of the solution into the *node-software* (cf. par. 2.1).

Our implementation of the Query Store consists of two distinct software elements:

- an overlay to (and embedded into) the existing VAMDC *node software*, thus independent from any specific database.
- a set of centralized asynchronous web-services, which may be seen as a smart log-service. In what follows we will call this element *Query-Store service.*[8]

Concerning the data versioning and time-stamping, we have two different mechanisms:

- a coarse-grained one: a modification of any publicly available data at a given *data-node* induces an increment in the version of the data-node. We have indeed a mechanism for informing that something has changed on a given *data-node*: in other words, we know that the result of an identical query may be different from one version to the other.
- a fined-grained one: based on the introduction of the *Version element* into the *VAMDC-XSAMS* standard, as described in (Zwölf et al. (2016)). The information contained into the *Version element* indicates which data have changed between two different *data-node* versions.

The Query Store is built over the coarse-grained mechanism.

3.1 The functioning of the Query Store

For extracting data from VAMDC, the users may query directly a given known *data-node* or use one of the centralized query-clients (e.g. the VAMDC portal, https://portal.vamdc.eu). In the latter case, the centralized client software asks the *registries* what are the *data-nodes* able to answer and dispatches to them the query. Any centralized client acts as a relay. This is complitely transparent from the *data-node* perspective and a *data-node* acts in the same way regardless the source of the query it is serving: when a *data-node* receives a query:

- it generates a unique *query-token* (this can be seen as a session token associated to the incoming query);
- it answers the query by producing the *VAMDC-XSAMS* output file, which is returned to the user together with the generated *query-token*. This token is copied both in the header of the answer and in the output file;
- it notifies to a specific notification service of the *Query-Store service* the *query-token*, the content of the query, the version of the node and the version of the standards used for formatting the output. It is worth noting that this process is not blocking and has no impact on the existing infrastructure whatsoever: the data extraction process is not slowed down and, if the Query Store

---

[8] We implemented these web-services using the Java Servlet technology©. The source code is released with a 'Creative Commons 4 (By, Nd, Nc)' license on GitHub: https://github.com/VAMDC/QueryStore.

cannot be reached the user will still receive the *VAMDC-XSAMS* output file.

When the *Query-Store service* receives a notification from the *data node*, it stores the received information and reduces the query to a standard form (using the VAMDC SQL-comparator library,[9] cf. remark 1 for a discussion) and it checks if a semantically identical query has already been submitted to the same *data-node*, having the same node version and working with the same version of the standards:

· If there is no such a query, the *Query-Store service* attributes a unique UUID and a timestamp to the new query, downloads the data, i.e. the *VAMDC-XSAMS* output file from the data-node and processes this file in order to extract the bibliographic information (each *VAMDC-XSAMS* file produced by the VAMDC infrastructure includes the references to the articles used for compiling the data) as well as metadata. The relevant metadata are stored and associated with the generated UUID. These metadata are kept permanently, while the downloaded XSAMS data are kept for an arbitrary time and then deleted (cf. remark 2 for a discussion).
· If such a query is already stored in the *Query-Store service*, the new couple (query time-stamp, query token) is added to the lists of the other time-stamps already associated with the query.

The *Query-Store service* permanently keeps the mapping between the UUID and the set of *query-tokens*[10] assigned to a given query. This information is kept for different reasons:

· statistics: it is interesting for database owner to know which queries are submitted and how many times a given query is re-submitted. This information is used for reporting to our founders and stakeholders.
· coherence of the human-interface: a user who has just re-submitted a query which was played for the first time long time ago by another user, may believe that there is some bug on the system if only the original timestamp is returned. By returning all the re-execution timestamp we avoid any ambiguity.
· troubleshooting and technical support: if something goes wrong on the Query-Store service before it issued the final UUID, we may use the token for identifying the query who generated the problem. Indeed the token is the first element generated into the query-notification pipeline.

During the query-submission phase the user has no direct interaction with the *Query-Store service* (as we explained before, the *data-node* that answers the query, notifies directly its action to the Query-Store). When the user receive the data from the *data-node* he/she has no information about the UUID the *Query-Store service* assigned to his/her query. The user may recover the final UUID assigned to his/her query by sending the query token to a specific service endpoint of the Query-Store (plus further optional information, e.g. the user e-mail and/or ORCID, information about the used client, etc...). This mechanism is implemented into the VAMDC-client software and its complexity is transparent to the scientific-user.

The functioning of the Query-Store is asynchronous. This was a mandatory constraint in order to avoid slowing down the VAMDC-infrastructure with a central bottleneck service. Indeed the *Query-Store service* response time could be slowed down if a huge number of queries comes in at the same time. Moreover computing the uniqueness of an incoming query may take some time if a very large number of queries is already stored. The asynchronous architecture solves these problems. A direct technical consequence of this asynchronous implementation is the combined generation of the associated tokens: the *query-token* and the *query-UUID*.

The unique identifier assigned to each query is resolvable, and is both human and machine actionable. The associated landing page provides the metadata associated with the query, as well as the access to the queried data. **Figure 1** represents a screen capture of the human-oriented landing page, whereas **Figure 3** represents the data model behind the *Query-Store service*.

As we can see in **Figure 3**, some personal information is stored into the *Query-Store service* (mainly the query submitted by the user). This information is kept only for internal purpose and in order to get a better user experience (cf. par 4.1). Because of this personal information and in the context of the European General Data Protection Regulation, we are registrating the *Query-Store service* with the CNIL (French National Agency regulating Data Protection).[11] All public interfaces of the *Query-Store service* are completely

---

[9] https://github.com/VAMDC/VamdcSqlRequestComparator.
[10] A query may be re-executed several times. Each execution has a different *query-token*.
[11] Since the Paris Observatory hosts the *Query-Store service* and is the legal representative of the VAMDC Consortium, we are subject to French law.

Figure 1: Screen capture of the human-oriented landing page for a given query. The "Data source", "Data source version", "XSAMS version", "Query" fields indicate respectively which *data-node* produced the result, the version of the *data-node*, the version of the standards when the query was processed, and the content of the query. The "Query identifier" is the UUID assigned by the *Query-Store service* to this query. The "Query Result downloaded on" list recall when this query was submitted (or re-submitted) and the "References" list contains the bibliographic references used for compiling the output file. For these, it is possible to switch between a tabular or a BibTex view (cf. figure 2). Finally a link gives access to the output file produced by the *data-node* while answering the query.

de-identified by virtually cutting the link "submitted-by" between the *Submission* and the *Author* classes: the queries contained into the *Query-Store service* may be browsed online in their anonymized form at the web-page: https://cite.vamdc.eu.

Remark 1 Comparing the sematic equivalence of two SQL queries is a problem which admits neither analytical nor close solution. This implies that 'false negative' may exist in the *VAMDC-SQL comparator* library which is built using the ANTLR parser.[12] Indeed if two queries are considered identical, they are actually identical; however in some minority cases, two semantically identical queries may be considered different.

Remark 2 Most of the queries processed by the VAMDC e-infrastructure are not used in a published work. It is therefore neither possible nor reasonable, to store the *VAMDC-XSAMS* data produced by all the queries for a very long term; data deletion is an operational requirement. The deletion mechanisms works as follow: the *XSAMS*-data produced by the *data-node* and stored on the *Query-Store service* are deleted only if the last (re-)execution of the query dates more than an arbitrary defined duration (5 years in our implementation). In other words, the data are not deleted if the query is too old but if the last query invocation is too old. Some specific queries (e.g. the queries associated with the Hidrogen H-α emission line, at a wavelength
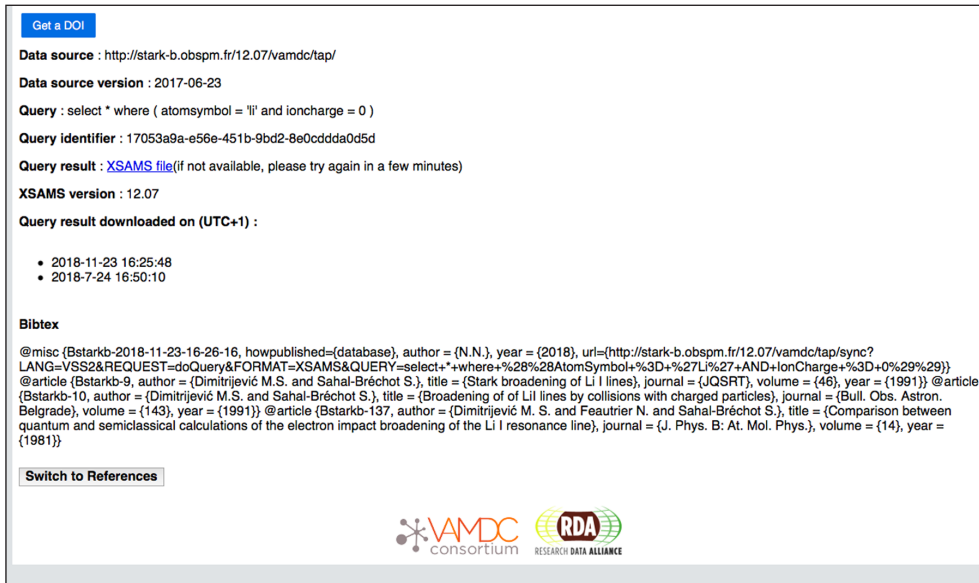
---

[12] https://www.antlr.org/.

Figure 2: Screen capture of the human-oriented landing page for a given query where a BibTex view is chosen for displaying the references. By clicking on the "Switch to References" button, one goes back to the display of figure 1.
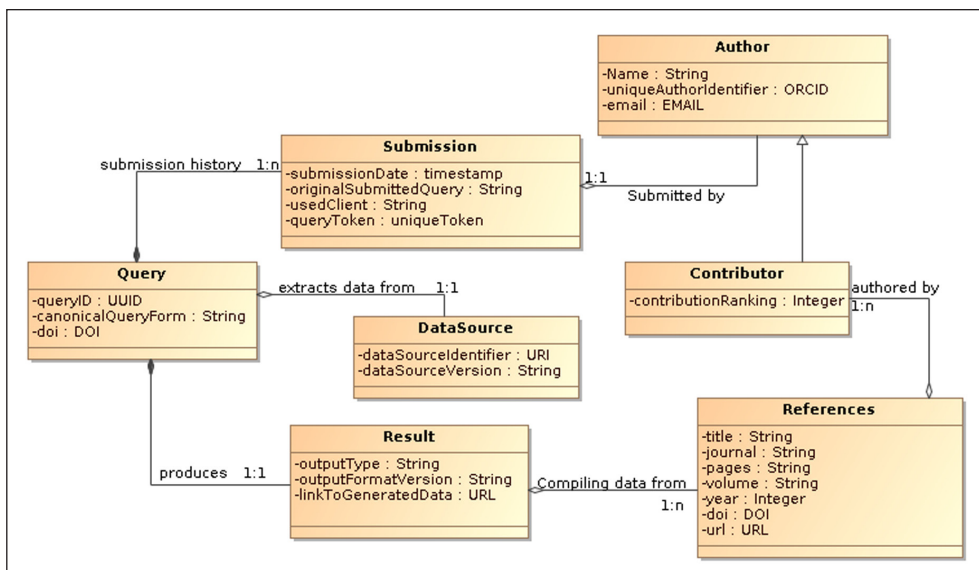

Figure 3: UML graphical representation of the data model used for organizing the metadata available in the *Query-Store service*.

of 656.28 nm, are commonly used for solar observations or detecting Hidrogen in space nebulae) may be very old, but re-executed daily. It is worth noting that only the *XSAMS*-data associated with the queries may deleted. All the other information (Data Source name and version, the query syntax and identifier, re-execution timestamps, bibliographic references) are kempt permanently.

The *XSAMS*-data associated with queries which have been assigned a DOI (i.e. which have been uploaded to Zenodo, cf. section 4.1) will never been deleted, regardless of their age.

4 Implementing Scholix for the Query Store

The Scholix recommendation is not implemented directly on the Query Store, but is a consequence of the interlinking between the *Query-Store service* and the Zenodo open science repository.[13]

---

[13] https://zenodo.org.

4.1 Interlinking the Query Store with Zenodo

As we highlighted in the remark 2, most of the queries are not cited by published works, and after an arbitrary time the underlying data are deleted from the *Query-Store service*. On the other hand, we have queries generating data used and cited in published works. A lifetime access must be provided to these data. The interconnection with Zenodo provides the Query Store with Scholix functionalities and with lifetime access to the query-generated data.

The link between the *Query-Store service* and Zenodo is implemented using, on the Query Store side, the Zenodo public REST API.[14] As we see in **Figure 1**, when a user uses the *Query-Store service* for displaying the information related to a given query, a button "Get a DOI" is displayed (if the query has not already been assigned a DOI). By clicking on this button, the user may trig the Zenodo registration process:[15] the file associated with the query is uploaded to Zenodo using the "Data Set" upload type and all the query-associated metadata are copied to corresponding Zenodo fields. In particular:

· the author of the upload is set to "VAMDC consortium";
· the title and the description are generated automatically starting from the query itself, the node producing the data and the query execution context (timestamp, token,…);
· the license chosen for the data being uploaded is "CC4 By", with open access;
· the bibliographic references extracted from the data-file by the *Query-Store service* while it processed the query (cf. paragraph 3.1), are copied into the "References" fields. The authors of these references also populates the "Contributors" fields;
· a reverse link, pointing from Zenodo to the *Query-Store service* query-entry, is introduced by putting into the field "Related Identifier – Is identical to" the resolvable persistent identifier of the query on the Query Store side (cf. remark 4 for a discussion about the relevance of this link).

When the upload process finishes successfully, Zenodo provides the *Query-Store service* with a DOI and with a deposition identifier, that the Query Store curators may use further for administrating the upload on the Zenodo side. These two identifiers are stored on the *Query-Store service* and associated to the query. The deposition identifier is never returned to the users. When a user displays a query which has already been copied to Zenodo, the button "Get a DOI" is replaced by a DOI badge (cf. **Figure 4**). By clicking on this badge, the corresponding Zenodo record is displayed on the user screen (cf. **Figure 5**). This page also contains the instructions for citing this query-record. Different export fromats are supported (cf. **Figure 6**): **Figure 7** gives an example of the Bibtex citation format. A citation for the query record we used for our example is (Consortium VAMDC (2018)).

Moreover, the data deletion mechanism described in 2 is suspended for all the queries associated with a DOI (in other words, the underlying data are kept permanently on the Query Store side as well).

Since Zenodo is indexed in OpenAIRE,[16] and since the latter implements Scholix through its Data-Literature Interlinking Service,[17] all the VAMDC queries registered by the Query Store in Zenodo are included in those infrastructures. Therefore when some data extracted from VAMDC are cited (in papers and/or other datasets) through the DOI obtained by the couple (Query Store/Zenodo), the authors of the works referenced by the VAMDC data receive credits automatically.

What has been described above is typical of the interoperability virtuous circle: if a system A implements some interoperability protocols and a system B implements some other ones, than a wrapping between A and B will disclose to A the interoperability capabilities of B. One could say that the interoperability-capabilities propagation speed is greater than the interoperability-protocols adoption speed.

Remark 3 The upload-to-Zenodo process was described in this paragraph from a human point of view. In our implementation this process may also be completely machine actionated. Indeed, the computer architecture of the services described through this paper relies on a set of REST services. A user, or a computer program, may interact with these services by sending parameters (using GET and/or POST methods) to specific endpoints. All these services respond by providing JSON formatted output which may be automatically parsed. The Graphical Web User interface we presented in **Figures 1, 2** and **4** are part of a lightweight html5 layer for interacting with and formatting output from these REST services.

---

[14] https://developers.zenodo.org.
[15] Automatic checks are implemented in order to avoid to register twice a given query.
[16] https://www.openaire.eu.
[17] https://scholexplorer.openaire.eu/index.html.

Figure 4: When a DOI is assigned, the "Get a DOI" button (cf. figures 1 or 2) is replaced by the DOI badge.

Remark 4 As Zenodo is an open repository (any person who has, e.g., a Github or an ORCID account may upload his/her productions to Zenodo), one has to pay the greatest attention to the provenance and to the scientific relevance of the works shared through this repository. In this context, the reverse link pointing from Zenodo to the Query-Store query-entry (see. **Figure 5**) gives users quality and provenance assurance on the shared datasets: the reverse link states that the data come from a well-known and documented database.

5 On-going and further works
The VAMDC funders and stakeholders regularly ask us to report on the outcomes of their investments, to track and demonstrate they have been used efficently. In this context the VAMDC Query-Store may play a double role: on one hand it may increase the impact of VAMDC (cf. section 5.1) and on the other it constitutes a fine-grained reporting tool (cf. section 5.2).

5.1 The Query-Store impact
As we underlined in (Moreau et al. (2018)), from the start of the VAMDC project in 2009, one of our goal has been to increase the citation impact of data producers. Indeed we find that the current status of citing spectroscopic data is to cite the database. It should be stressed that atomic and molecular data require months to be either measured or calculated, and therefore it is a loss of visibility and recognition that only databases be cited in users' papers. We believe that the Query Store coupled to the VAMDC portal now allow this flaw to be overcome, even if additional refinements need to be carried out: on January 2018 we have started the deployement of the Query-Store data citation capabilities in the production environment. Currently these are deployed over a third of the *Data-nodes* of the VAMDC infrastructure. Since January 2018 the *Query-Store service* received ~2000 queries. From these, ~180 unique queries have been identified. The link between the Query-Store and Zenodo have been added in May 2018. Since then ~10 queries received a DOI.

Figure 5: Partial screen-shot of the Zenodo-record landing page: The mentioned query token is the one generated by the node serving the query (cf. section 3.1). The set of references are those provided by the *Query-Store service* during the submission phase. One can also see on the right side the *reverse link* (Related Identifiers) pointing to the original query record on the VAMDC side.

This paper is the first peer-reviewed work where the technical details of the VAMDC Query Store are described, whereas (Moreau et al. (2018)) is the first article where the atomic/molecular science-aspects linked with the Query Store are discussed: at this point in time we cannot be sufficiently objective for evaluating how the usage of the VAMDC infrastructure has been altered by the implementation of the Query Store.

We would like the Query-Store to boost the usage of VAMDC. For that reason, we have recently started collaborating with the main Astronomy and Physics Journal editors so that they may have their paper-submission workflows adapted for interacting directly with the VAMDC Query Store. Our goal is to make the Query Store indispensable for all the author publishing papers citing atomic or molecular data. During the submission phase, the author may put references to data using the DOIs assigned by the Query Store, as it is already the case for papers. We are working with editors for achieving this integration. All the actors will obtain benefits: VAMDC will increase its impact and its usage, editors will gain an efficient tool for data-paper linking and data producers/providers will benefit from the automatic citation mechanisms. From the earlier discussion with editors, we have identified some improvement targets, described in section 5.3.

5.2 Refining the level of autorhip

As described in on (FAIR Data (2018)) (Recommendation 6), data practitioners should facilitate the inclusion of a wide range of indicators for the assessment of the scientific and technical contributions to data-related activities: provision of data infrastructure ans services should be recognized and rewarded accordingly. In order to be able to measure all the contributions (together with the specific role of each contributor) we are planning to extend the range of metadata to be sent to Zenodo. Indeed Zenodo adopts the DataCite Metadata Schema for the Publication and Citation of Research Data (DataCite Metadata Working Group (2017)). This schema is very rich and contains several optional parameterswe we would like to exploit: we think that the fields "Contributor – Data Collector", "Contributor – Data Curator", "Contributor – Data Manager" are very valuable since, by filling those fields, we may mention and acknowledge with bibliometric credits the work of people involved in VAMDC data-infrastructure maintenance and curation. Nowadays this technical work is anonymous and mostly invisible for the scientific final users of VAMDC.
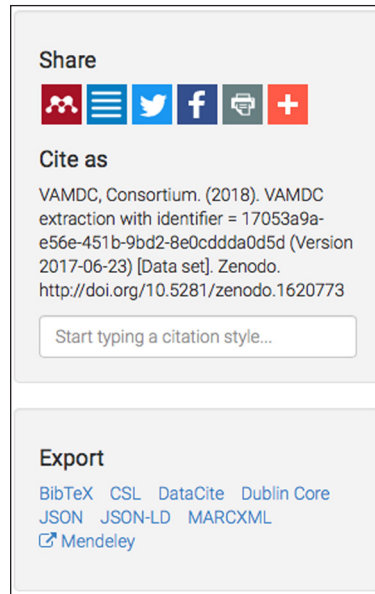
Figure 6: Partial screen-shot of the Zenodo-record landing page: this part of the screen displays the instruction for citing the current query-record.
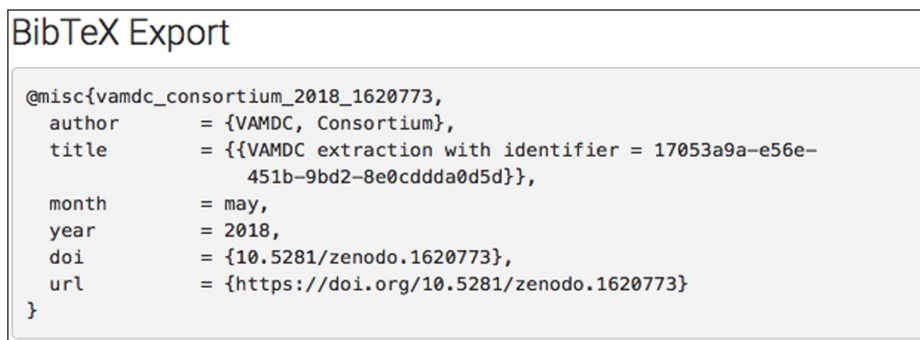

Figure 7: Bibtex format to be used for citing the query-record of this example.

The VAMDC registry (cf. 2.1) already contains the names of the scientific and technical maintainers of each *data-node*, however, there do not exist machine actionable mechanisms for extracting this information from the current version of the registry. We are developing such a service: while registering to Zenodo a query processed by a given *data-node*, the *Query-Store service* will extract -directly and on the fly- from the registries the information about the scientific and technical curator of the *data-node*. The Zenodo fields concerning the "Data Contributors" will be populated accordingly.

5.3 Clustering queries

In order to enhance the user experience, we would like to provide new services for clustering a set of queries and assign to the cluster a DOI. The service we are designing relies on user-authentication and authorization. An authenticated user will be able to:

- create a new query-Cluster. He/she will automatically be the first author of the freshly created cluster;
- add other contributors to an existing query-Cluster (he/she is first-authoring). The new authors may be added by their identifier (typically their ORCID) and by specifying their contribution rank to the cluster (e.g. 2nd author, 3rd author,…);
- add/remove queries to a cluster he/she is co-authoring. Only the queries whose data have not yet been deleted may be added to a cluster). The automatic data deletion mechanism (cf. par. 4.1) will also be blocked for the queries belonging to a query cluster;
- publish to Zenodo the query cluster he/she is first-authoring. This will assign a DOI to the cluster.

In **Figure 8** we represent all the metadata (together with their structure) attached to a given query cluster. As we can see on **Figure 8** we have three different levels of authorship:

- the author contributing to the cluster. An author may contribute to the cluster, without having any submitted query;
- the author who processed a given Query attached to the cluster (this author is always part of the authors of the cluster);
- the author who wrote a paper referenced by the result of a Query.

All these three levels of authorship are important and, through the Zenodo metadata schema (cf. par. 4.1), will receive credits when a given Query Cluster is cited through its DOI.

While implementing these additional features to the VAMDC Query Store, we will pay great attention to follow the RDA recommendations on Data Collections.[18]

6 Concluding remarks

Through this paper we exposed how the new RDA data citation paradigms have been implemented on the VAMDC distributed e-infrastructure and how we succeeded in removing the technical barriers linked with the automatic data-citation and with the delegation of credits for VAMDC-extracted data. However, the success of a technical solution does not only depend on its intrinsic quality, but also on its level of adoption by the user-community: we are focusing our efforts:

- on increasing the impact of the described citation services through community awareness-raising and training around these new tools, as we suggested in (Moreau et al. (2018)).
- Working with editors for integrating the VAMDC Query Store in the paper submission workflows (for paper citing atomic and molecular data, cf. section 5.1).
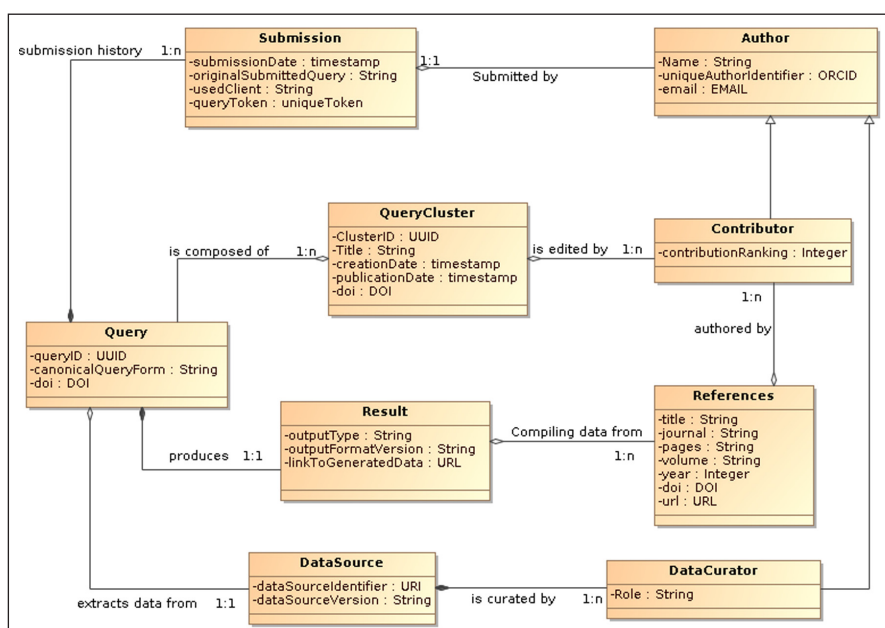

Figure 8: Graphical representation of the meta data associated with a query cluster: Each cluster is created and/or modified by specific contributors (a first author, a second author, etc.). Each author may add to the clusters the Queries he/she performed (provided the query related data are still present and not deleted from the system, cf. remark 2). Each Query produces a result by extracting data from a specific data source. Each result has a list of references, i.e. the list of all the publication used for compiling the result. The unique identifier of each Cluster is resolvable and the associated landing page will contain all the cluster-associated metadata, together with access to the underlying data (i.e. all the data coming from the extraction performed by the queries composing the cluster).

---

[18] https://rd-alliance.org/group/research-data-collections-wg/outcomes/rda-research-data-collections-wg-recommendations.

We are focusing our efforts in this collaboration with editors because we believe this is a key action to consolidate the VAMDC position as a leading infrastructure for sharing atomic and molecular data.

Acknowledgements

Competing Interests

The authors have no competing interests to declare.

References

Asmi, A, Rauber, A, Pröll, S and **van Uytvanck, D.** 2016. Citing Dynamic Data – Research Data Alliance working group recommendations. In: *EGU General Assembly Conference Abstracts*, *volume 18*, *EGU General Assembly Conference Abstracts*, EPSC2016–7456. April 2016.

Bell, G, Hey, T and **Szalay, A.** 2009. Beyond the Data Deluge. *Science*, 323(5919): 1297–1298. ISSN 0036-8075. URL: https://science.sciencemag.org/content/323/5919/1297. DOI: https://doi.org/10.1126/science.1170411

Benson, K, Plante, R, Auden, E, Graham, M, Benson, K, Plante, R, Greene, G, Hill, M, Linde, T, Morris, D, O'Mullane, W, Rixon, G, Stébé, A and **Andrews, K.** 2009. IVOA Registry Interfaces Version 1.0. *IVOA Recommendation* 04 November 2009.

Burton, A, Fenner, M, Haak, W and **Manghi, P.** November 2017. Scholix Metadata Schema for Exchange of Scholarly Communication Links. DOI: https://doi.org/10.5281/zenodo.1120265

Consortium VAMDC. VAMDC extraction with identifier = 17053a9ae56e-451b-9bd2-8e0cddda0d5d, May 2018. DOI: https://doi.org/10.5281/zenodo.1620773

DataCite Metadata Working Group. 2017. DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.1. DataCite e.V. https://schema.datacite.org/meta/kernel-4.1/index.html. DOI: https://doi.org/10.5438/0014

Dowler, P, Rixon, G and **Tody, D.** 2010. Table Access Protocol Version 1.0. *IVOA Recommendation* 27 March 2010.

Dubernet, ML, Antony, B, Ba, Y-A, Babikov, Y, Bartschat, K, Boudon, V, Braams, B, Chung, H-K, Daniel, F, Delahaye, F, Del Zanna, G, de Urquijo, J, Dimitrijevic, M, Domaracka, A, Doronin, M, Drouin, B, Endres, C, Fazliev, A, Gagarin, S, Gordon, I, Gratier, P, Heiter, U, Hill, C, Jevremovic, D, Joblin, C, Karsprzak, A, Krishnakumar, E, Leto, G, Loboda, PA, Louge, T, Maclot, S, Marinkovic, B, Markwick Kemper, A, Marquart, T, Mason, H, Mason, N, Mendoza, C, Mihajlov, A, Millar, T, Moreau, N, Mulas, G, Pakhomov, Y, Palmeri, P, Pancheshnyi, S, Perevalov, VI, Piskunov, N, Postler, J, Quinet, EL, Sánchez, PQ, Ralchenko, Y, Rhee, Y-J, Rixon, G, Rothman, L, Roueff, E, Ryabchikova, T, Sahal-Brechot, S, Scheier, P, Schlemmer, S, Schmitt, B, Stempels, E, Tashkun, S, Tennyson, J, Tyuterev, V, Vujcic, V, Wakelam, V, Walton, N, Zatsarinny, O, Zeippen, C and **Zwölf, CM.** 2016. The Virtual Atomic and Molecular Data Centre (VAMDC) Consortium. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(7). DOI: https://doi.org/10.1088/0953-4075/49/7/074003

Dubernet, ML, Boudon, V, Culhane, JL, Dimitrijevic, MS, Fazliev, AZ, Joblin, C, Kupka, F, Leto, G, Le Sidaner, P, Loboda, PA, Mason, HE, Mason, NJ, Mendoza, C, Mulas, G, Millar, TJ, Nuñez, LA, Perevalov, VI, Piskunov, N, Ralchenko, Y, Rixon, G, Rothman, LS, Roueff, E, Ryabchikova, TA, Ryabtsev, A, Sahal-Bréchot, S, Schmitt, B, Schlemmer, S, Tennyson, J, Tyuterev, VG, Walton, NA, Wakelam, V and **Zeippen, CJ.** 2010. Virtual atomic and molecular data centre. *J. Quant. Spectrosc. & Rad. Transfer*, 111: 2151–2159. Oct, 2010. DOI: https://doi.org/10.1016/j.jqsrt.2010.05.004

European Commission Expert Group on FAIR Data. 2018. Turning FAIR into reality. *Final report and Action Plan*. URL: http://www.codata.org/news/254/62/Turning-FAIR-Data-into-Reality-Report-and-Action-Plan-Consultation-until-5-August.

Hey, T, Tansley, S and **Tolle, K.** 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. *Microsoft Research*, October. ISBN: 978-0-9825442-0-4. URL: https://www.microsoft.com/en-us/research/publication/fourth-paradigm-data-intensive-scientific-discovery/.

Moreau, N, Zwölf, CM, Ba, Y-A, Richard, C, Boudon, V and **Dubernet, M-L.** 2018. The VAMDC Portal as a major vector of atomic and molecular data citation. *Galaxies*, pages galaxies-326995.

Regandell, S, Marquart, T and **Piskunov, N.** March 2018. Inside a VAMDC data node – putting standards into practical software. *Physica Scripta*, 93(3). DOI: https://doi.org/10.1088/1402-4896/aaa268

VAMDC Consortium. 2012. VAMDC SQL Subset, version 2. *VAMDC standard.* http://vamdc.eu/documents/standards/queryLanguage/vss2.html.

Walton, N. 2004. Meeting the User Science Challenge for a Virtual Universe. In: *Toward An International Virtual Observatory: Proceedings Of The Eso-esa-nasa-nsf Conference Held At Garching*, 188. Germany, 10–14 June 2002. DOI: https://doi.org/10.1007/10857598_29

Wilkinson, MD, Caselli, P, Pon, A, Belloche, A and **André, P.** March 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3. Online. DOI: https://doi.org/10.1038/sdata.2016.18

Wittenburg, P, Hellström, M, Zwölf, C-M, Abroshan, H, Asmi, A, Di Bernardo, G, Couvreur, D, Gaizer, T, Holub, P, Hooft, R, Häggström, I, Kohler, M, Koureas, D, Kuchinke, W, Milanesi, L, Padfield, J, Rosato, A, Staiger, C, van Uytvanck, D and **Weigel, T.** December 2017. Persistent identifiers: Consolidated assertions. Status of November, 2017. DOI: https://doi.org/10.5281/zenodo.1116189

Zwölf, CM, Dubernet, M-L, Ba, Y-A and **Moreau, N.** May 2014. Experience and feedbacks from the sustainability for the virtual atomic and molecular data centre E-infrastructure. In: *IST-Africa Conference Proceedings*, 1–9. DOI: https://doi.org/10.1109/ISTAFRICA.2014.6880621

Zwölf, CM, Moreau, N and **Dubernet, M-L.** September 2016. New model for datasets citation and extraction reproducibility in VAMDC. *Journal of Molecular Spectroscopy*, 327: 122–137. DOI: https://doi.org/10.1016/j.jms.2016.04.009