The VAMDC Portal as a Major Enabler of Atomic and Molecular Data Citation

Nicolas Moreau, Carlo-Maria Zwolf, Yaye-Awa Ba, Cyril Richard, Vincent Boudon, Marie-Lise Dubernet

galaxies

MDPI

Article

The VAMDC Portal as a Major Enabler of Atomic and Molecular Data Citation

Nicolas Moreau [1], Carlo-Maria Zwolf [1], Yaye-Awa Ba [1], Cyril Richard [2], Vincent Boudon [2] and Marie-Lise Dubernet [1,*,†]

[1] Laboratoire d'Étude du Rayonnement et de la Matiére en Astrophysique, Observatoire de Paris, UMR CNRS 8112, UPMC, 5 Place Jules Janssen, 92195 Meudon CEDEX, France; nicolas.moreau@obspm.fr (N.M.); carlo-maria.zwolf@obspm.fr (C.-M.Z.); yaye-awa.ba@obspm.fr (Y.-A.B.)

[2] Laboratoire Interdisciplinaire Carnot de Bourgogne, UMR 6303 CNRS-Univ. Bourgogne Franche-Comté, 9 Avenue Alain Savary, BP 47 870, F-21078 DIJON CEDEX, France; Cyril.Richard@u-bourgogne.fr (C.R.); Vincent.Boudon@u-bourgogne.fr (V.B.)

* Correspondence: marie-lise.dubernet@obspm.fr; Tel.: +33-01-4507-7570

† Current Address: Observatoire Aquitain des Sciences de l'Univers, Bât. B18N, Allée Geoffroy Saint Hilaire, CS 50023, 33615 PESSAC CEDEX, France.

Received: 25 June 2018; Accepted: 25 September 2018; Published: 3 October 2018

Abstract: VAMDC bridged the gap between atomic and molecular (A&M) producers and users through providing an interoperable e-infrastructure connecting A&M databases, as well as tools to extract and manipulate those data. The current paper highlights the usage of the VAMDC Portal, recalls how data citation is implemented within VAMDC and provides insights about usage of VAMDC that will increase the impact factor of A&M producers and will offer a more reliable citation of A&M datasets included in application fields.

Keywords: atom; molecules; database; citation

1. Introduction: State-of-the-Art

The "Virtual Atomic and Molecular Data Centre Consortium" (VAMDC Consortium, http://www.vamdc.eu) [1] is a worldwide consortium which federates Atomic and Molecular databases through an e-science infrastructure and an organisation to support this activity (http://www.vamdc.org/structure/how-to-join-us/). About 90% of the inter-connected databases handle data that are used for the interpretation of astronomical spectra and for the modeling in media of many fields of astrophysics. Other application fields include atmospheric physics, plasmas, fusion, radiation damage.

The current VAMDC e-infrastructure interconnects about 30 atomic and molecular databases that cover atomic and molecular spectroscopy and processes. VAMDC offers a common entry point to all encorporated databases through the VAMDC portal (http://portal.vamdc.eu) and VAMDC develops also standalone tools in order to retrieve and handle the data, the SPECTCOL tool [2,3] is an example (http://www.vamdc.eu/software). VAMDC provides also software [4] and support in order to include new databases within the VAMDC e-infrastructure. One feature of VAMDC e-infrastructure is the constrained environment for the description of data, in particular the *VAMDC-XSAMS*, a standard XML file (XML Schema for Atomic Molecular and Solid Data)[1], and other standardized protocols (http://www.vamdc.eu/standards) that ensure a higher quality for the distribution of data. *VAMDC-XSAMS* is an evolved version of the XSAMS schema presented by [5] and it should be used together with

[1] https://standards.vamdc.eu/dataModel/vamdcxsams/index.html#vamdcxsamslanguage-index

the description of molecular quantum numbers provided by what we call the case-by-case schema[2]. Our recent publication [1] provides details about *VAMDC-XSAMS* and about the main databases included in the VAMDC e-infrastructure.

By 2016 the VAMDC Consortium started to collaborate with the Research Data Alliance (RDA)[3] and to work in its Data Citation Working Group[4]. Indeed the VAMDC Consortium intended to find a way for users to cite the datasets that the infrastructure provides. The RDA Data Citation Working Group provided the researchers and data centres communities with a recommendation to identify and cite dynamic data [6]: 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 will store all the identified and time-stamped queries, together with relevant metadata and availability to recover the data as it existed at the time when a given query was executed.

Recently VAMDC, commissioned by the Research Data Alliance, has implemented the recommendations [6] of the RDA data citation group. Indeed VAMDC provides an interesting science use case of a typical distributed infrastracture with geographically distributed databases, registries (meaning "yellow pages") and access tools. Within this context a first work has been done on provenance of datasets [7], meaning that versioning and data-timestamping has been included in the VAMDC- XSAMS schema. The second work, for which RDA provided funding to the VAMDC Consortium, has implemented the concept of *Query Store* which impacts both the node software [4], the registries and has been implemented in the VAMDC Portal (http://portal.vamdc.eu). The technical description of the Query Store that stores timestamped queries submitted to the VAMDC infrastructure, can be found in [8].

The current paper aims at presenting the *VAMDC Portal* coupled to the *Query Store*. We re-call the key features of the VAMDC e-infrastructure, we show how the VAMDC Portal provides the users with the ability to access the Query Store and thus create DOI for their datasets, and finally we discuss the potential impact of the VAMDC citation features for data users through presenting different science use cases.

2. Results: VAMDC Portal and Query Store

The *VAMDC portal* uses the VAMDC standards and technological developments, it provides a seamless access to the inter-connected VAMDC databases. Through this unique interface as displayed on Figure 1, a user can query any database member of the VAMDC infrastructure, and can retrieve data in the common shared file format *VAMDC-XSAMS*[5].

In order to be visible from the portal, as from any VAMDC tool or user, each database must provide a VAMDC compatible access thanks to *the node software* (as described in Section 2.1).

2.1. Connecting Heterogeneous Databases into the VAMDC Interoperable Infrastructure—The Role of the Node Software

The e-infrastructure connects 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 is successful in removing the bottleneck between data producers and the wide body of A&M data users. 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* [4] transforms an autonomous database into a VAMDC federated database, called *data-node*. Each *data-node* accepts queries submitted in

---

2 http://www.vamdc.eu/documents/cbc-1.0/
3 https://www.rd-alliance.org
4 https://www.rd-alliance.org/groups/data-citation-wg.html
5 http://www.vamdc.eu/standards

a standard grammar (cf. Section 2.3) and provides output formatted into *VAMDC-XSAMS*[6]. Then each database must be registered into a central repository called a *registry* that provides a standardized application programming interface (API) to explore its content and to discover the available VAMDC resources (cf. Section 2.2).

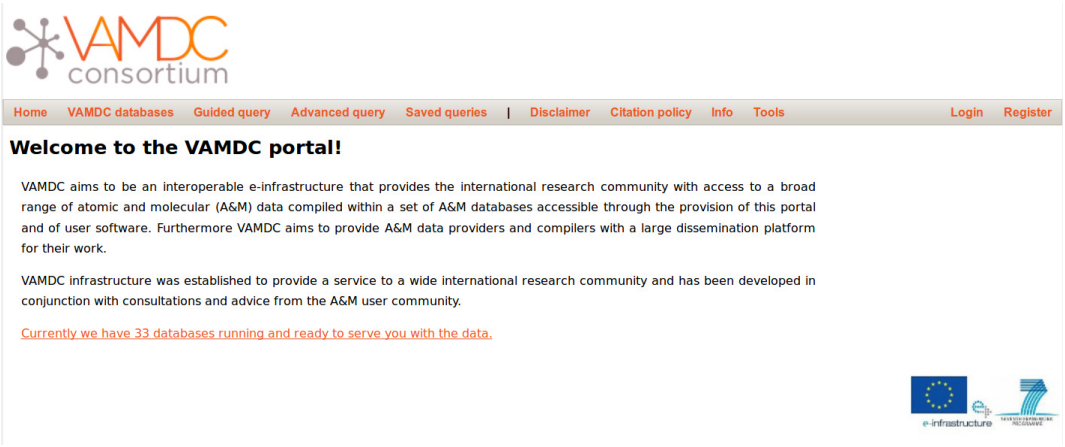


Figure 1. VAMDC Portal welcome page: Userssee they may access data using a simplified or advanced interface, they may save queries, have access to tools, and must accept the citation policy and the condition of the disclaimer. Information are provided to users in real-time about the number of databases interconnected by the VAMDC infrastructure.

2.2. Registry

The VAMDC *registry* is located at http://registry.vamdc.eu/registry-12.07/main/index.jsp. It is based on the work of the Astrogrid project[7] that was the UK's Virtual Observatory development project from 2001 to 2010 [9]. They developed a registry whose interface is based on the International Virtual Observatory Alliance standard [10]. Thus any user or service can easily know how to write queries to find the services. Each of the VAMDC service is registered in the VAMDC *registry* as a *VOResource* [11] to describe its metadata (service name, address, query parameters, type of data). To simplify the access to this registry, the VAMDC consortium provides a java library[8], that is used by the VAMDC portal, among other applications.

2.3. Query Language

Another key element used by the VAMDC portal, and central to the VAMDC infrastructure, is the *query language* which we chose to be a subset of SQL, called *VAMDC QSL Subset*[9], and that allows the user to query multiple databases simultaneously. This query language is understood by the *data access protocol VAMDC-TAP*[10] implemented on each database. The protocol accepts such queries and return files in the *VAMDC-XSAMS* format. *VAMDC-TAP* is a simplified version of the *TAP protocol* from the IVOA [12]. *VAMDC-TAP* exposes only one table, simplifying the query by removing the join part. To achieve this result, a *dictionary* has been defined[11]. Each quantity that any database can potentially return is defined in the *dictionary* with a given keyword. Then the *node software* can do the mapping

6 http://www.vamdc.eu/standards
7 http://www.astrogrid.org/
8 http://www.vamdc.eu/documents/standards/registry/queryingForResources.html
9 https://standards.vamdc.eu/queryLanguage/index.html
10 http://www.vamdc.org/documents/standards/dataAccessProtocol/vamdctap.html
11 http://dictionary.vamdc.eu

between this quantity and an actual column in the database. For example a request returning all the data related to the helium atom is written:

Select all where AtomSymbol = 'he'

2.4. Species Database

In order to provide an efficient search environment, being able to search by species name is a key element. However, as necessary as it might be, it is has been proven to be a complex matter.

Looking for an atom is simple as it is efficiently identified with its name or its symbol. For a molecule though, things become more complex. A molecule's formula can be written in several ways, it can also be searched by its stoichiometric formula or by one of the many standards or classification that have been developed (SMILE, Inchi, CAS number ...).

To overcome this difficulty, the VAMDC infrastructure created a centralized chemical species repository, called the *species database*. Updated daily, it contains the list of all the species in each of the VAMDC databases. Each species is identified uniquely by an *InChiKey*, an identifier generated from an *InChi* description[12]. In the *species database* each InChiKey is associated to the different ways to identify a species, e.g., their chemical names, formula, stoechiometic formula, CAS number. By adding a REST API[13] and a web graphical interface to this species database (https://species.vamdc.eu), we provide a versatile tool to explore the species content of the atomic and molecular VAMDC connected databases.

In addition to relying on this REST API of the species database, the VAMDC portal provides both an auto-completion possibility and an isotopologues discovery feature. So it becomes possible to specify very precisely which species is the most relevant to your search.

2.5. Portal Search Interface

There are two search interfaces available as displayed on Figure 1. Both of them are graphical tools that build *VAMDC-TAP* requests, and thus that mask the complexity of the internal language to the users.

The first graphical interface follows a step by step approach as displayed on Figure 2. Each time a user chooses an option, a limited set of new options appears. It is particularly recommended for the people discovering the portal as the user is guided throughout the selection process.

The second interface, which is called "advanced", lets the users choose by themselves their search criteria. They can compose their own request by combining "Species", "Processes" and "Environment" characteristics, as it is displayed on the left hand side of Figure 3. Each time the content of the query is updated, the list of databases that can answer is displayed in green on the right of Figure 3.

2.6. Portal Results Display, Query Store and Data Citation

Once a request has been completely parameterized by the user, it is converted into a *VAMDC-TAP process*, and is sent to each database that knows about the sent *dictionary keywords* (c.f. Query section). The results are displayed in a table where each line is the answer from a database, as shown on Figure 4.

For each database we provide a summary of the returned content, i.e., the number of species, states, of processes that can be described as collisional transitions, radiative or non-radiative transitions following the *VAMDC-XSAMS schema*. Then the user can click on the "XSAMS File" button in order to download the data in a *VAMDC-XSAMS* format. The *VAMDC-XSAMS* format can be uploaded in other tools such as the SPECTCOL tool [2,3] for collisional and radiative transition, or can be viewed by any editor.

---

[12] https://iupac.org/who-we-are/divisions/division-details/inchi/

[13] A RESTful API is a method of allowing communication between a web-based client and server that employs representational state transfer (REST) constraints

The second column of the table, called "View data", provides a list of transformation options for the *VAMDC-XSAMS file*. The display options can convert the XML contained in the *VAMDC-XSAMS* file to another format, such as an HTML page with export functions, an HITRAN format [13] for molecular data, a bibtex generated from references attached to the data. With those processors the user can use directly the data without having to use any other tools.

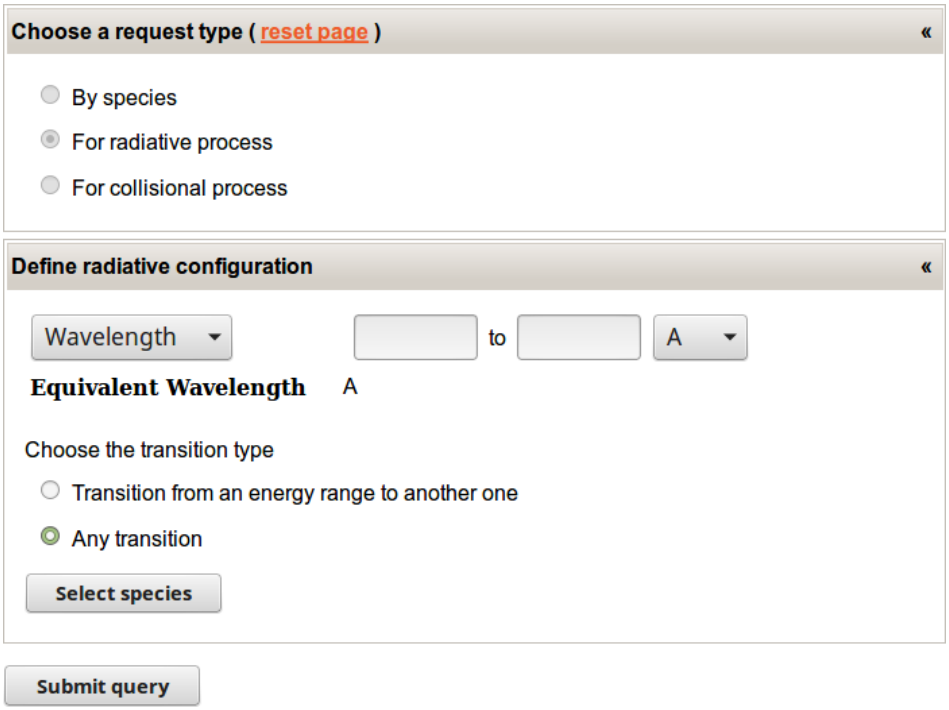


Figure 2. Query process in the guided mode.

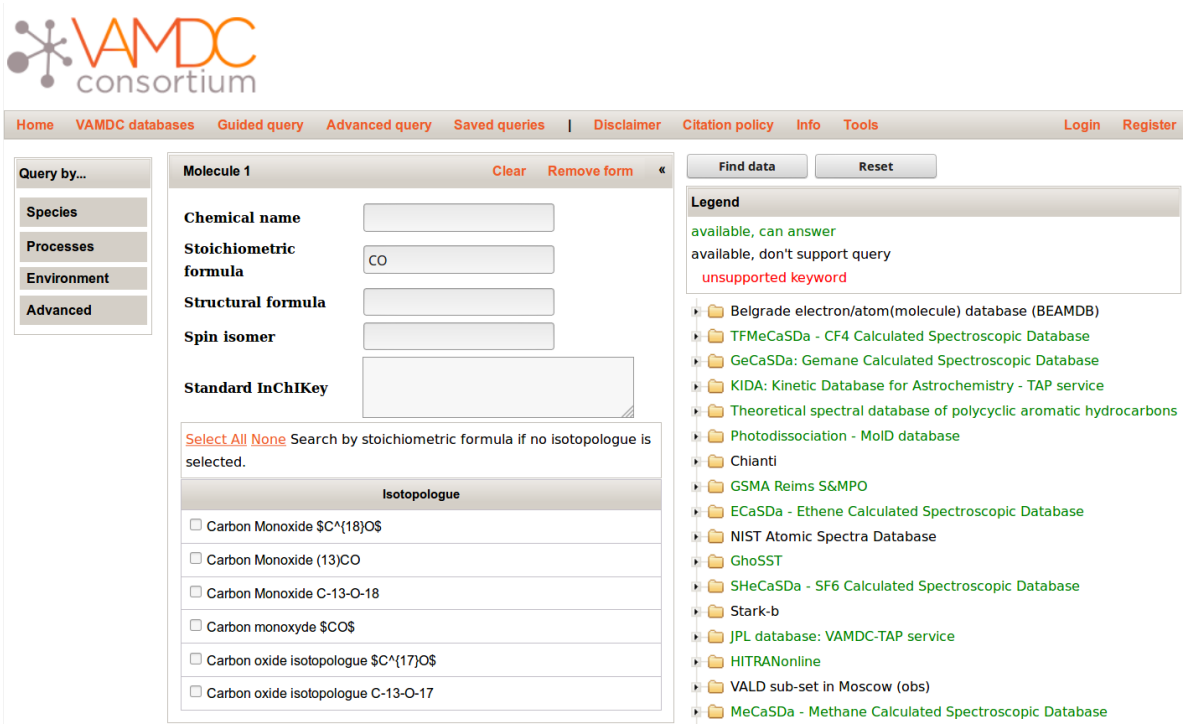


Figure 3. Query Process in the advanced mode. This query is based on searching the stoechiometric formula. Then a choice of isotopologues is retrieved from the *species database* and proposed for selection.

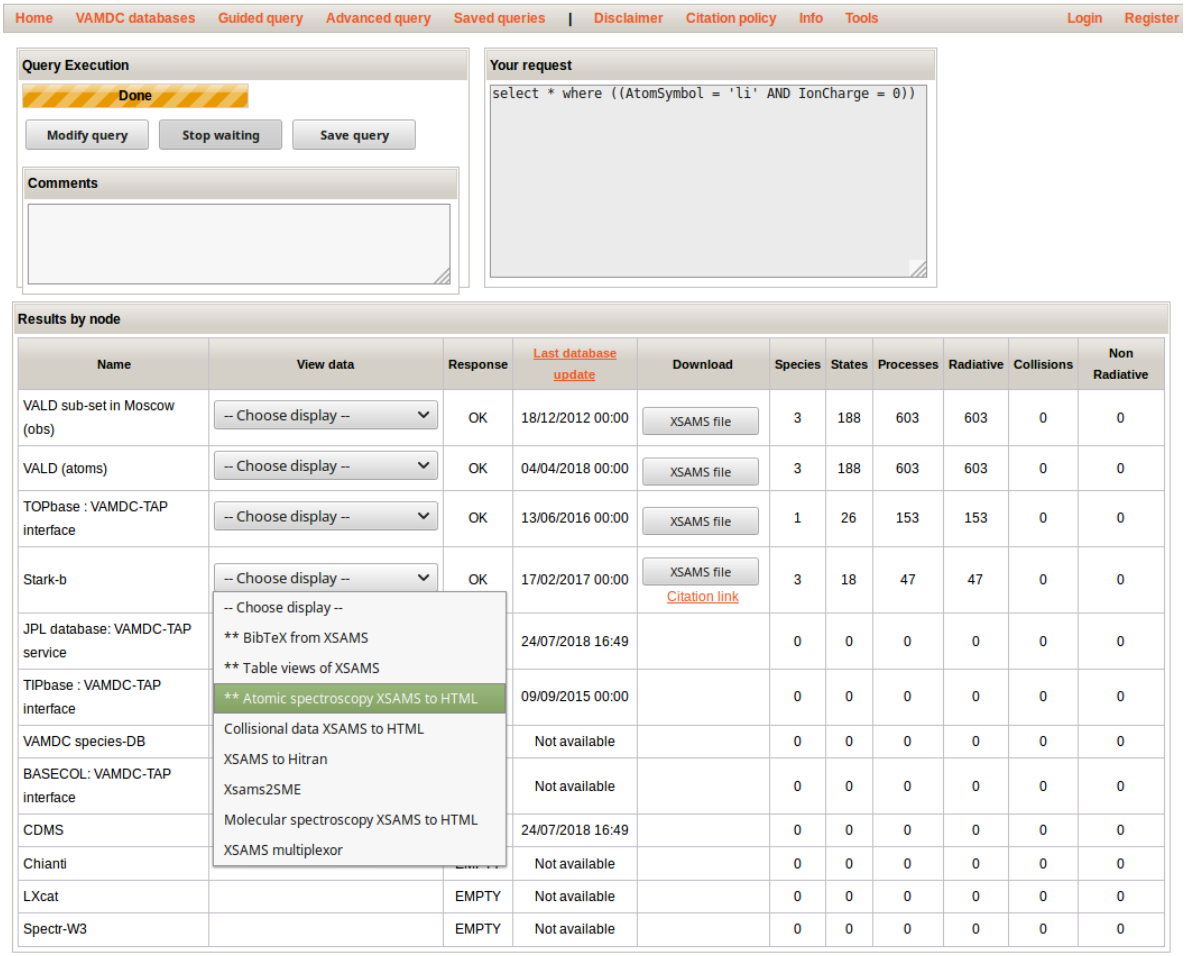
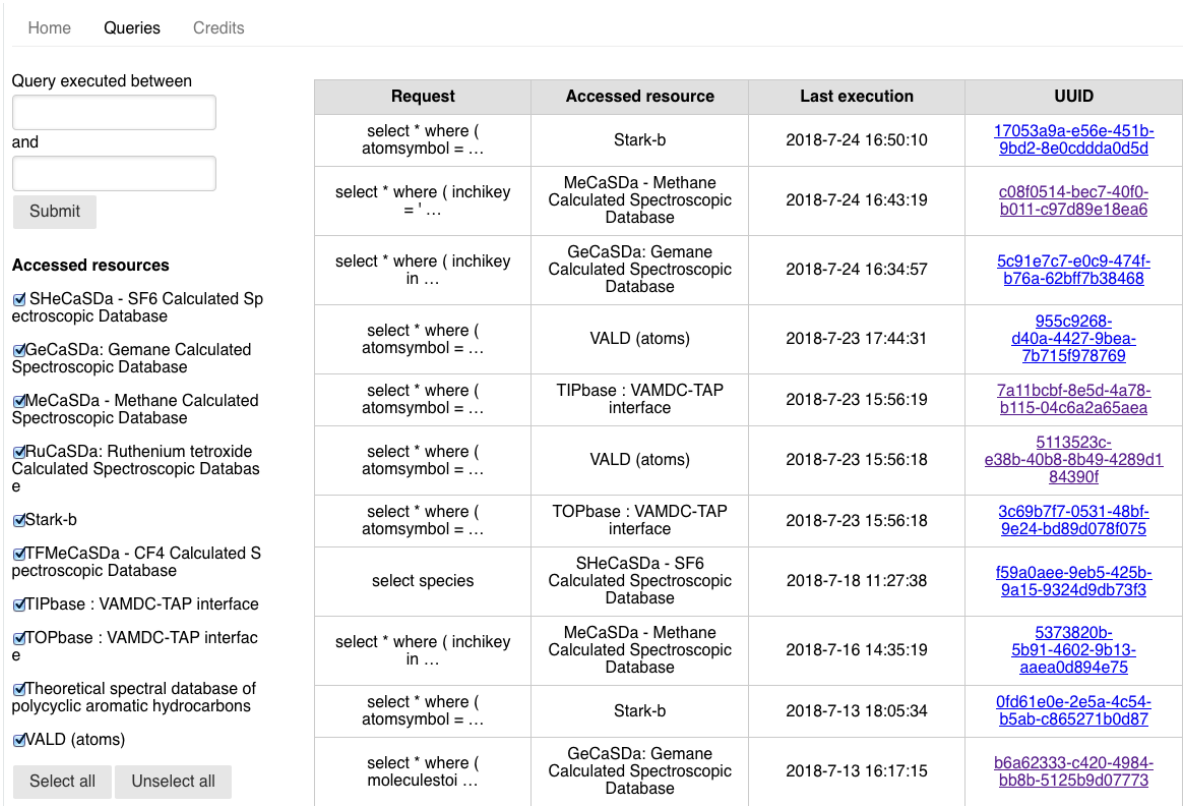


Figure 4. Result page once a query is sent. The query is shown in the "Your request" box and corresponds to a search of Li I.

Finally under the "XSAMS file" button, a "Citation link" is displayed. This citation link leads to the *Query Store* as displayed on Figure 5. The *node software* [4] has been upgraded to be the bridge between the client software used by the final user, here the *VAMDC Portal* and the *Query-Store*. The *node software* generates a *token* that is notified to the *Query Store* along with the request, and this *token* is also returned to the VAMDC Portal. More technical details can be found in [8]. The VAMDC Portal uses the *token* to get the persistent identifier of the request from the *Query Store*. Once it has received it, it displays the "Citation link" in the result page of Figure 4, this link goes to the landing page at https://cite.vamdc.eu/persistentId, as represented in Figure 6. This landing page (Figure 6) is the typical human readable Query-Store landing page that a user reaches when he resolves the persistent identifier associated with a given query.



Request	Accessed resource	Last execution	UUID
select * where (atomsymbol = …	Stark-b	2018-7-24 16:50:10	17053a9a-e56e-451b-9bd2-8e0cddda0d5d
select * where (inchikey = ' …	MeCaSDa - Methane Calculated Spectroscopic Database	2018-7-24 16:43:19	c08f0514-bec7-40f0-b011-c97d89e18ea6
select * where (inchikey in …	GeCaSDa: Gemane Calculated Spectroscopic Database	2018-7-24 16:34:57	5c91e7c7-e0c9-474f-b76a-62bff7b38468
select * where (atomsymbol = …	VALD (atoms)	2018-7-23 17:44:31	955c9268-d40a-4427-9bea-7b715f978769
select * where (atomsymbol = …	TIPbase : VAMDC-TAP interface	2018-7-23 15:56:19	7a11bcbf-8e5d-4a78-b115-04c6a2a65aea
select * where (atomsymbol = …	VALD (atoms)	2018-7-23 15:56:18	5113523c-e38b-40b8-8b49-4289d184390f
select * where (atomsymbol = …	TOPbase : VAMDC-TAP interface	2018-7-23 15:56:18	3c69b7f7-0531-48bf-9e24-bd89d078f075
select species	SHeCaSDa - SF6 Calculated Spectroscopic Database	2018-7-18 11:27:38	f59a0aee-9eb5-425b-9a15-9324d9db73f3
select * where (inchikey in …	MeCaSDa - Methane Calculated Spectroscopic Database	2018-7-16 14:35:19	5373820b-5b91-4602-9b13-aaea0d894e75
select * where (atomsymbol = …	Stark-b	2018-7-13 18:05:34	0fd61e0e-2e5a-4c54-b5ab-c865271b0d87
select * where (moleculestoi …	GeCaSDa: Gemane Calculated Spectroscopic Database	2018-7-13 16:17:15	b6a62333-c420-4984-bb8b-5125b9d07773

Figure 5. Query-Store web interface: the VAMDC databases implementing the *Query Store* feature at the time of this publication are listed into the *Accessed Resources* list.

The landing page stores the persistent identifier (named "Query identifier" in Figure 6) associated with the query, the query itself, the name and version of the node answering the query, the dataset produced by the *node* while processing the query, together with the bibliographic references extracted from the dataset. This bibliographic information displays the references cited in the the *Source* element of the *VAMDC-XSAMS* file, and those references might be associated to any type of data included in the dataset. For a finer-grained understanding of how these references span the different elements of the dataset, one must investigate the *VAMDC-XSAMS* file generated by the query. In the future we will improve the management of this fine granularity.

Currently the citation feature is implemented on about a third of the VAMDC connected databases as displayed in real-time on Figure 5 below "Accessed resources". Once a node has implemented the *Query Store* feature, any requests to that node are registered in the *Query Store*. The *Query Store* content may be directly explored at the url "https://cite.vamdc.eu". By clicking on the *Query* tab, the results represented in Figure 5 are displayed: the user may filter the results by date or by database (*Accessed resources*) and search for a persistent identifier to which is attached the corresponding landing page. With the *Query store*, the VAMDC infrastructure has a way to remember queries permanently. The user can then use the uniquely generated and persistent link to view the detail of his query at a later date and even download the data again.

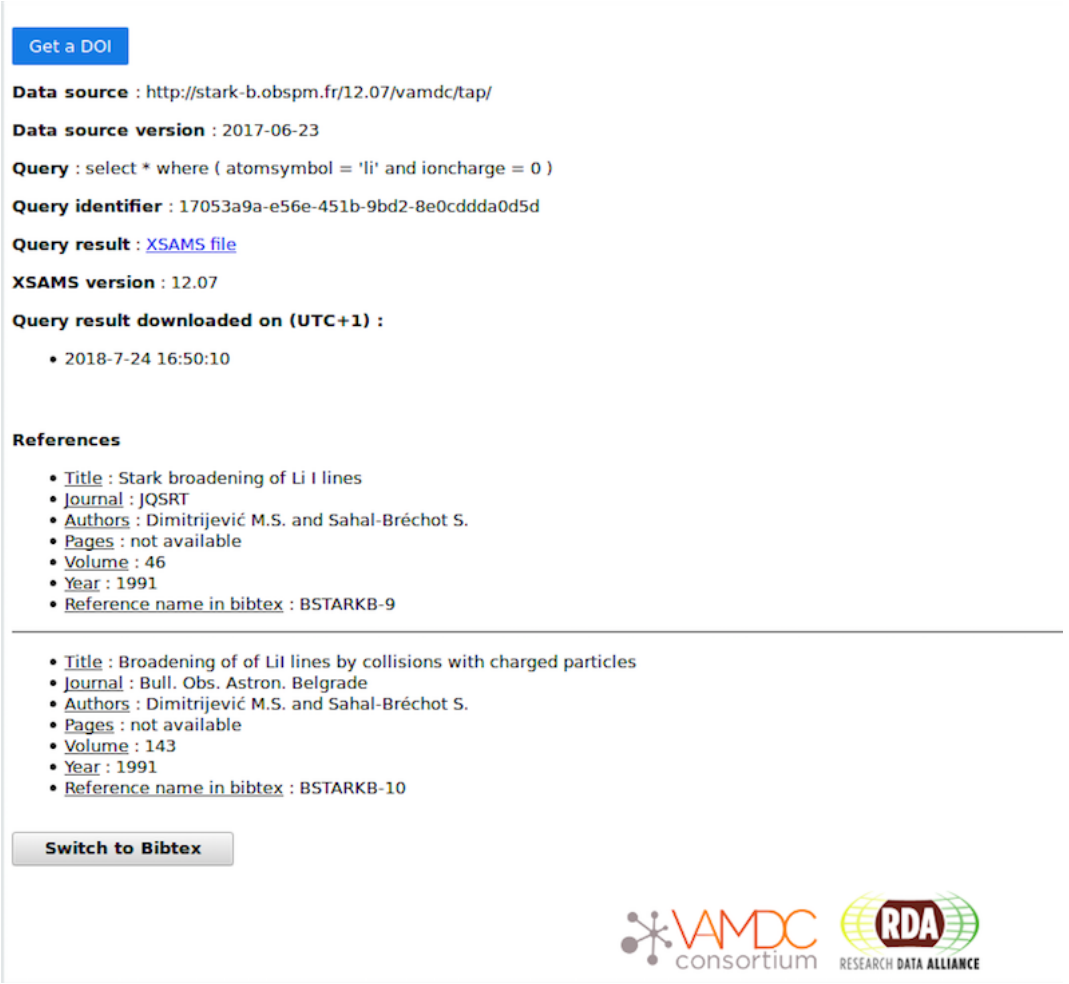


Figure 6. Human readable Query-Store landing page, obtained while resolving the persistent identifier associated with a Query. In the present case the persistent identifier is: "17053a9a-e56e-451b-9bd2-8e0cddda0d5d". The database is STARK-B [14].

2.7. Beyond the Actual Data Citation Processes Used by Editors: The Scholix Initiative and the Query Store

Nowadays editors have essentially two basic mechanisms for linking articles to data repositories. One concerns entity linking, where the journal picks up a unique identifier or code in an article and establishes a deep link to the underlying data deposited elsewhere. The other one is banner linking, where an automated query is sent each time a new article is published to the external database in order to check whether there is data available for this article. We explain the limitation of these approaches with the following examples:

- Let A be an article referencing a dataset D. If the data repository containing D has no ab initio idea that the article A exists, it has to search through all the Internet and through all the services of the existing journals if a paper citing D exists. The data trafic worldwidely generated by this approach is enormous, if we consider that there exists a lot of data centers containing thousands of datasets and that each datacenter, for each data-set it contains, will ask the same question to the same journals (currently approximatively 8M articles are registered into the different editor online services).
- Let us consider a datacenter where a dataset D' contains references to an article A'. How may the publisher of A' know that D' exists?

These two examples show that the data-citation model currently adopted by the editors is not sustainable and does not meet the data-driven science community. The *Scholix* initiative [15] succeded in establishing a high level interoperability framework for exchanging information about the links between scholarly literature and data. It has been adopted by existing hubs or global aggregators of data-literature link information such as DataCite, CrossRef, OpenAIRE, EMBL-EBI, together with Eslevier and Springer editors.

The Scholix recommendation is not implemented directly on the *Query Store*, but is a consequence of the interlinking between the *Query-Store* and the Zenodo open science repository[14]. The link between the *Query Store* and Zenodo is implemented using the Zenodo public REST API[15] in the Query Store software. The landing page of Figure 5 displays a button "Get a DOI" if the query has not already been assigned a DOI. By clicking on this button, the user triggers the Zenodo registration process: 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. When the upload process finishes successfully, Zenodo provides the Query-Store with a DOI which is stored in the Query-Store and associated to the query. When a user displays a landing page/query which has already been copied to Zenodo, the button "Get a DOI" is replaced by a DOI badge.

Zenodo is indexed in OpenAIRE[16], OpenAIRE implements Scholix through its Data-Litterature Interlinking Service[17], therefore all the VAMDC queries registered by the Query Store in Zenodo are included in those infrastructures. 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 credit automatically. Indeed the scholarly links harvested from the *Data-Litterature Interlinking Service*[18] flow to the *hubs* implementing Scholix (e.g., CrossRef and DataCite). In addition members from the SAO/NASA Astrophysics Data System (ADS) are active members of the RDA-Scholix Working Group and are working at implementing Scholix in ADS; therefore one implemented credits will flow automatically to the ADS system.

3. Discussion: Science Use Cases for Data Citation Impact

Therefore the VAMDC Portal provides access to the landing pages identified with Unique Identifiers. As all queries are stored in the Query Store for a period of time, users can find the landing page at a later stage using the Unique Identifier that they can store. We present below several science use cases where the implemented citation features can be used to identify the datasets that are kept in their final analysis.

3.1. Prospective Use of VAMDC in Spectroscopy

3.1.1. Analysis of Astrophysical Spectra

An example of usage is the need to query atomic or molecular lines on a given frequency or wavelength range for Local Thermodynamic Equilibrium analysis. The VAMDC infrastructure facilitates greatly the obtaining of large quantities of data to which an equally large quantity of references are attached. The usual strategy of users is to cite the databases that have been queried, but rarely the original authors of the papers. Before VAMDC it was very painful to collect all the references, so it was understandable. With VAMDC the references are readily available, but the number of citation pages would outnumbered the content of the paper. We believe that this problem is now solved with our system. Users can query any databases, they can download the data files, use them in various

---

[14] https://zenodo.org
[15] http://developers.zenodo.org
[16] https://www.openaire.eu
[17] http://scholexplorer.openaire.eu/index.html#/api
[18] the harvested information may be explored at http://scholexplorer.openaire.eu/index.html

applications, and prior to publication of their spectra analysis, he will decide which spectroscopic data have been the most relevant to their study. For those data they will assign a DOI to each dataset through the Query Store and will cite the DOIs in their publication. As explained above (see Section 2.6) the authors of the works referenced by the VAMDC data receive credit automatically.

3.1.2. A Scientific Use Case: Intercomparison of Databases

We provide below a scientific use case that could be attractive to atmospheric physicists and planetologists. These communities usually use the HITRAN 160-character file format [13] as an input for their radiative transfer or atmospheric modeling codes. Thus, the original *VAMDC-XSAMS* output format produced by the *data-nodes*[19] is not directly nor easily useable by the different users. Within this context, we developed a new tool for converting to the HITRAN file-format any *VAMDC-XSAMS* file produced by any *data-node* containing molecular data; this application follows the *VAMDC-XSAMS consumer protocol*[20] standard of VAMDC. In order to reach this new level of interoperability, several databases had to perform some adjustments to their contents. For instance, the HITRAN intensity unit (at 296 K) was added to the *VAMDC-XSAMS* standard; the CDMS [16–18] and JPL [18,19] databases added a new field for this intensity unit in their *VAMDC-XSAMS* processor.

As already mentioned, the HITRAN conversion (see Figure 4) is available for all the molecular databases already integrated into the VAMDC infrastructure (e.g., HITRAN [20], CDMS [16–18], JPL [18,19], MeCaSDa [21,22], ...): this allows quick and direct comparisons between them. To achieve this aim, a graphical chart can be produced by loading two HITRAN output files, either coming from the VAMDC portal thanks to the above-mentioned converter or from the HITRAN On line service[21]. The following on-line tool[22] might be used for that purpose. Such combination of tools may help data producers to check the consistency of their data and to point out database content differences. For instance, some databases include experimental, fitted or calculated ab initio data only, or mixings between these different sources, etc. Also, some databases include isotopic abundance factors, some other not.

Figure 7 gives an example of database comparison between the HITRAN and JPL databases: such comparison would have been very cumbersome in the past. We mention that we intend to provide the same tools for the GEISA [23] file format as soon as possible.

If VAMDC already provided interoperable access to all its *data-nodes*, the described HITRAN processor provides the community with an easy tool for comparing and crossmatching data coming from heterogeneous molecular databases. Molecular-scientists may particularly welcome these new features. As the potential adoption of this tool is wide, embedding the citation feature into the VAMDC portal is very important for data producers and providers: for a given analysis, the relevant and interesting data can be uniquely identified through the VAMDC-citation feature. The combined usage of the processor for visualisation and of the data-citation feature of the VAMDC portal provides a new paragdim for carrying out analysis of spectra.

3.2. Prospective Use of VAMDC in Numerical Codes Packages: Example of Cloudy and PDR Code

Cloudy is a non-local thermodynamic equilibrium spectral synthesis and plasma simulation code designed to simulate astrophysical environments and predict their spectra. A recent effort [24] has been to move Cloudy's atomic and molecular data into external databases. They use external databases such as CHIANTI[23] and LAMDA[24]. For some ions they indicate using data from version 7.1.4 of the

---

[19] We remember (cf. Section 2.1) that we call *data-node* a database which joined the interoperable VAMDC e-infrastructure
[20] https://standards.vamdc.eu/#xsams-processor-service
[21] http://hitran.org
[22] http://www.vamdc.org/hitran-display/
[23] http://www.chiantidatabase.org/
[24] http://www.strw.leidenuniv.nl/~moldata/

CHIANTI database. For versioning of LAMDA it is indicated that they downloaded the LAMDA data on the 30 June 2015. For data citation they advise users to go to the website of the CHIANTI [25] and of the LAMDA [26] databases.

It should be noted that CHIANTI is part of VAMDC, when LAMDA is not. Nevertheless the collisional data of LAMDA are in the BASECOL [27] database. The BASECOL[25] database has a version attached to each of its dataset and can be queried via the VAMDC portal or with the SPECTCOL tool [2,3]. BASECOL will be referenced in the Query Store.

The Meudon PDR code[26] [28], that can be used to study the physics and chemistry of diffuse clouds, photodissociation regions (PDRs), dark clouds, is another example of numerical code where the atomic and molecular data are externalized.

By using VAMDC facilities to query data, it will be possible to uniquely identify the data in time and to cite the DOI in the code; again the producers of atomic and molecular data will be acknowledged through the pipeline that we have put together.

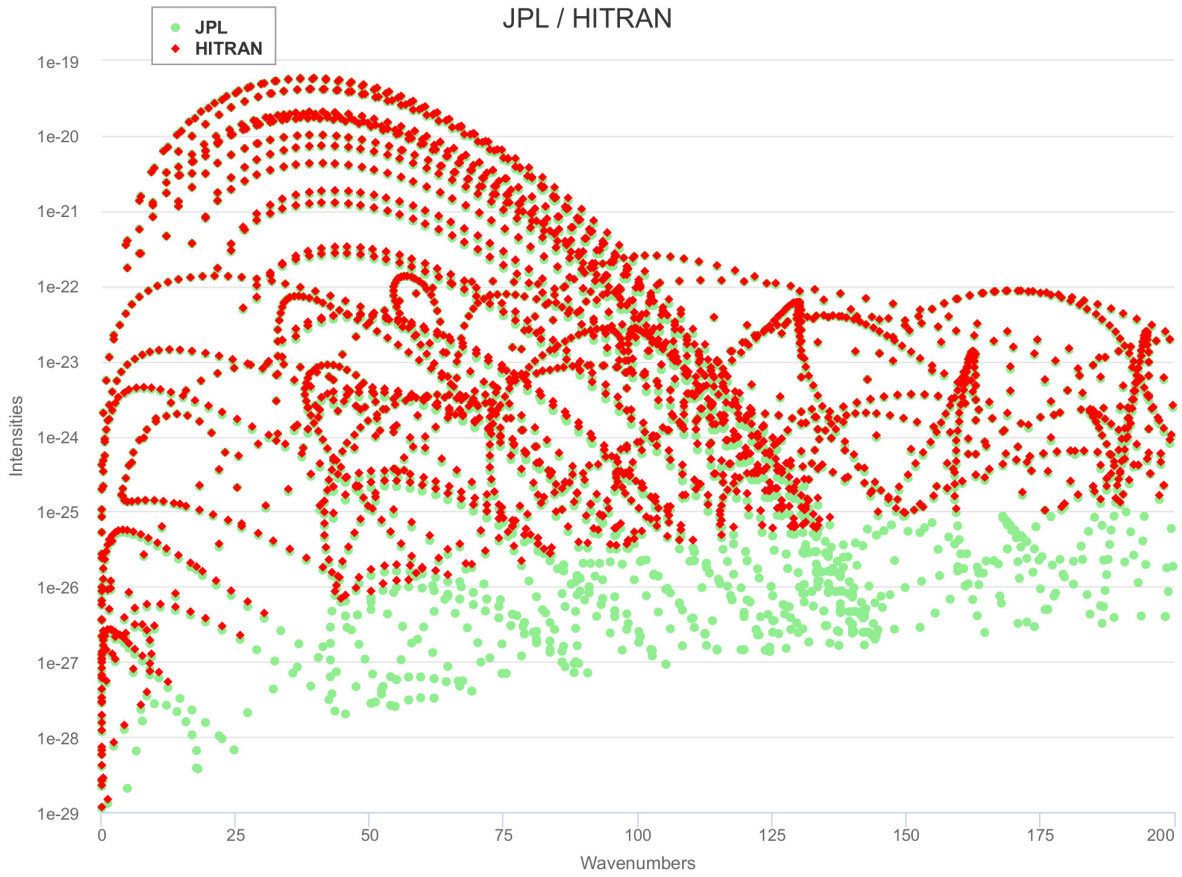


Figure 7. Comparison of formaldehyde (H_2CO) lines in the HITRAN and JPL databases, after extraction in HITRAN format from the VAMDC portal.

3.3. Prospective Use of VAMDC for Secondary Databases

The VAMDC e-infrastructure could be heavily used to produce secondary databases, for example to prepare the analysis of specific space missions or for other specific purposes. For example the Belgian repository of fundamental atomic data and stellar spectra (BRASS) [29] aims to provide the

[25] http://basecol.vamdc.org
[26] https://ism.obspm.fr/

largest systematic and homogeneous quality assessment of atomic data to date in terms of wavelength, atomic and stellar parameter coverage. To do so they retrieved atomic data from repositories and did cross-matching of data. They mention that the majority of repositories were retrieved via the VAMDC e-infrastructure and that they are grateful for the current efforts of the VAMDC team in homogenising the repositories as this has helped to expedite their comparisons and cross-match work. This is of course one of the main achievement of VAMDC: to facilitate retrieval and comparisons of data. The new citation feature will allow them to trace the data that they queried in the VAMDC repositories.

3.4. Conclusion and Future Work

From the start of the VAMDC project in 2009 one of our goals 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. For example a search of the ADS NASA system with the word "CDMS" in the text shows that only the CDMS database's [16–18] URL and/or reference are provided in the majority of the astrophysical papers. 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. This paper encourages users to explore the different tools and to provide the VAMDC collaboration feedbacks of usage in order to improve the system.

While writing this paper we have found another key interest of the *Query Store*, which is to have a better view of references attached to a given set of data. Of course references have been present in the *VAMDC-XSAMS* files and accessible from the visualisation tools for years. Nevertheless only with the *Query Store* did we figure out that references attached to some files were totally incoherent (we thank one of the referees for pointing out such incoherence). This is not related to the *Query Store*, but to the *VAMDC-XSAMS* files provided by the *node*. This shows that a large scientific survey must be carried out using the *Query Store* in order to improve the output *VAMDC-XSAMS* files. Another interesting remark of one of the referee was the absence of selection in the references list as the references might be attached to different quantities in the *VAMDC-XSAMS* files, and the user might not want to use all of them. This issue will be addressed in the future.

Currently only a few VAMDC connected databases have implemented the Query Store feature, Figure 6 shows the list of implemented databases at the time of publication. Future work includes the implementation of the citation feature on all the VAMDC connected databases by the end of 2019.

Some additional future technical work will be to group the queries so that only one DOI is assigned to several queries. Finally we are currently implementing the Query Store feature into SPECTCOL [2,3], another VAMDC tool that can query molecular spectroscopic and collisional databases, and then display and match the molecular data.

4. Materials and Methods

The VAMDC portal is fully operational and the query store has been implemented as described above. The users can now access the capability of the citation features, the number of implemented nodes will increase.

Author Contributions: Software, N.M., Y.-A.B., C.-M.Z., C.R.; Writing—Original Draft Preparation, N.M., C.-M.Z., V.B., C.R., M.-L.D.; Writing—Review & Editing, M.-L.D., C.-M.Z.; Funding Acquisition, M.-L.D. and C.-M.Z.

Funding: Support for VAMDC has been provided through the VAMDC and the SUP@VAMDC projects funded under the "Combination of Collaborative Projects and Coordination and Support Actions" Funding Scheme of The Seventh Framework Program. Call topic: INFRA-2008-1.2.2 and INFRA-2012 Scientific Data Infrastructure. Grant Agreement numbers: 239108 and 313284. We acknowledge support from Paris Astronomical Data Center of Paris Observatory. The Query Store was partially funded by the Research Data Alliance.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Dubernet, M.L.; Antony, B.; Ba, Y.A.; Babikov, Y.; Bartschat, K.; Boudon, V.; Braams, B.; Chung, H.K.; Daniel, F.; Delahaye, F.; et al. The Virtual Atomic and Molecular Data Centre (VAMDC) Consortium. *J. Phys. B At. Mol. Opt. Phys.* **2016**, *49*, 074003. [CrossRef]
2. Dubernet, M.L.; Nenadovic, L. *SPECTCOL: Spectroscopic and Collisional Data Retrieval*; Record ascl:1111.005; Astrophysics Source Code Library: Houghton, MD, USA, 2011; p. 11005.
3. Ba, Y.A.; Dubernet, M.L. SPECTCOL2018. In *To be to Submitted to: Molecular Astrophysics*; Elsevier: Amsterdam, The Nederthlands, 2018.
4. Regandell, S.; Marquart, T.; Piskunov, N. Inside a VAMDC data node—Putting standards into practical software. *Phys. Scr.* **2018**, *93*, 035001. [CrossRef]
5. Ralchenko, Y.; Clark, R.E.H.; Dubernet, M.L.; Gagarin, S.; Humbert, D.; Loboda, P.A.; Moreau, N.; Roueff, E.; Schultz, D.R. Development of new standards for exchange of atomic and molecular data. In Proceedings of the 6th International Conference on Atomic and Molecular Data and Their Applications, Beijing, China, 27–31 October 2008; pp. 207–216.
6. Asmi, A.; Rauber, A.; Pröll, S.; van Uytvanck, D. Citing Dynamic Data-Research Data Alliance working group recommendations. In Proceedings of the EGU General Assembly Conference Abstracts, Vienna, Austria, 17–22 April 2016; Volume 18.
7. Zwölf, C.M.; Moreau, N.; Dubernet, M.L. New model for datasets citation and extraction reproducibility in VAMDC. *J. Mol. Spectrosc.* **2016**, *327*, 122–137. [CrossRef]
8. Zwölf, C.M.; Moreau, N.; Ba, Y.A.; Dubernet, M.L. Implementing in the VAMDC the new paradigms for data citation from the Research Data Alliance. *Data Sci. J.* **2018**, Submitted.
9. Walton, N. Meeting the User Science Challenge for a Virtual Universe. In *Toward An International Virtual Observatory, Proceedings of the Eso-Esa-Nasa-Nsf Conference, Garching, Germany, 10–14 June 2002*; Springer: Berlin/Heidelberg, Germany, 2004; pp. 188–192.
10. Benson, K.; Plante, R.; Auden, E.; Graham, M.; Greene, G.; Hill, M.; Linde, T.; Morris, D.; O'Mullane, W.; Rixon, G.; et al. IVOA Registry Interfaces Version 1.0. IVOA Recommendation. 2009. Available online: http://xxx.lanl.gov/abs/1110.0513 (accessed on 4 November 2009).
11. Plante, R.; Benson, K.; Graham, M.; Greene, G.; Harrison, P.; Lemson, G.; Linde, T.; Rixon, G.; Stébé, A.; IVOA Registry Working Group. VOResource: An XML Encoding Schema for Resource Metadata Version 1.03. IVOA Recommendation. 2008. Available online: http://xxx.lanl.gov/abs/1110.0515c (accessed on 22 February 2008).
12. Dowler, P.; Rixon, G.; Tody, D. Table Access Protocol Version 1.0. IVOA Recommendation. 2010. Available online: http://xxx.lanl.gov/abs/1110.0497 (accessed on 27 March 2010).
13. Rothman, L.S.; Jacquemart, D.; Barbe, A.; Benner, D.C.; Birk, M.; Brown, L.R.; Carleer, M.R.; Chackerian, C., Jr.; Chance, K.; Coudert, L.H.; et al. The HITRAN 2004 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Trans.* **2005**, *96*, 139–204. [CrossRef]
14. Sahal-Bréchot, S.; Dimitrijević, M.S.; Moreau, N.; Ben Nessib, N. The STARK-B database VAMDC node: A repository for spectral line broadening and shifts due to collisions with charged particles. *Phys. Scr.* **2015**, *90*, 054008. [CrossRef]
15. Burton, A.; Fenner, M.; Haak, W.; Manghi, P. *Scholix Metadata Schema for Exchange of Scholarly Communication Links*; CERN: Geneva, Switzerland, 2017.
16. Endres, C.P.; Schlemmer, S.; Schilke, P.; Stutzki, J.; Mueller, H.S.P. The Cologne Database for Molecular Spectroscopy, CDMS, in the Virtual Atomic and Molecular Data Centre, VAMDC. *J. Mol. Spectrosc.* **2016**, *327*, 95–104. [CrossRef]
17. Müller, H.S.P.; Schlöder, F.; Stutzki, J.; Winnewisser, G. The Cologne Database for Molecular Spectroscopy, CDMS: A useful tool for astronomers and spectroscopists. *J. Mol. Struct.* **2005**, *742*, 215–227. [CrossRef]
18. Endres, C.; Schlemmer, S.; Drouin, B.; Pearson, J.; Müller, H.S.P.; Schilke, P.; Stutzki, J. Improved Infrastucture for Cdms and JPL Molecular Spectroscopy Catalogues. In Proceedings of the 69th International Symposium on Molecular Spectroscopy, Urbana, IN, USA, 16–20 June 2014.
19. Pickett, H.M.; Poynter, R.L.; Cohen, E.A.; Delitsky, M.L.; Pearson, J.C.; Muller, H.S.P. Submillimeter, millimeter, and microwave spectral line catalog. *J. Quant. Spectrosc. Rad. Trans.* **1998**, *60*, 883–890. [CrossRef]

20. Gordon, I.E.; Rothman, L.S.; Hill, C.; Kochanov, R.V.; Tan, Y.; Bernath, P.F.; Birk, M.; Boudon, V.; Campargue, A.; Chance, K.V.; et al. The HITRAN2016 molecular spectroscopic database. *J. Quant. Spectrosc. Radiat. Trans.* **2017**, *203*, 3–69. [CrossRef]
21. Ba, Y.A.; Wenger, C.; Surleau, R.; Boudon, V.; Rotger, M.; Daumont, L.; Bonhommeau, D.A.; Tyuterev, V.G.; Dubernet, M.L. MeCaSDa and ECaSDa: Methane and ethene calculated spectroscopic databases for the virtual atomic and molecular data centre. *J. Quant. Spectrosc. Radiat. Trans.* **2013**, *130*, 62–68. [CrossRef]
22. Amyay, B.; Gardez, A.; Georges, R.; Biennier, L.; Vander Auwera, J.; Richard, C.; Boudon, V. New investigation of the nu(3) C-H stretching region of (CH4)-C-12 through the analysis of high temperature infrared emission spectra. *J. Chem. Phys.* **2018**, *148*, 134306.
23. Jacquinet-Husson, N.; Armante, R.; Scott, N.A.; Chedin, A.; Crepeau, L.; Boutammine, C.; Bouhdaoui, A.; Crevoisier, C.; Capelle, V.; Boonne, C.; et al. The 2015 edition of the GEISA spectroscopic database. *J. Mol. Spectrosc.* **2016**, *327*, 31–72. [CrossRef]
24. Ferland, G.J.; Chatzikos, M.; Guzmán, F.; Lykins, M.L.; van Hoof, P.A.M.; Williams, R.J.R.; Abel, N.P.; Badnell, N.R.; Keenan, F.P.; Porter, R.L.; et al. The 2017 Release Cloudy. *arXiv* **2017**, *53*, arXiv:1705.10877.
25. Landi, E.; Del Zanna, G.; Young, P.R.; Dere, K.P.; Mason, H.E. CHIANTI—An Atomic Database for Emission Lines. XII. Version 7 of the Database. *Astrophys. J.* **2012**, *744*, 99. [CrossRef]
26. Schöier, F.L.; van der Tak, F.F.S.; van Dishoeck, E.F.; Black, J.H. An atomic and molecular database for analysis of submillimetre line observations. *Astron. Astrophys.* **2005**, *432*, 369–379. [CrossRef]
27. Dubernet, M.L.; Alexander, M.H.; Ba, Y.A.; Balakrishnan, N.; Balança, C.; Ceccarelli, C.; Cernicharo, J.; Daniel, F.; Dayou, F.; Doronin, M.; et al. BASECOL2012: A collisional database repository and web service within the Virtual Atomic and Molecular Data Centre (VAMDC). *Astron. Astrophys.* **2013**, *553*, A50. [CrossRef]
28. Le Petit, F.; Nehmé, C.; Le Bourlot, J.; Roueff, E. A Model for Atomic and Molecular Interstellar Gas: The Meudon PDR Code. *Astr. J. Sup.* **2006**, *164*, 506–529. [CrossRef]
29. Laverick, M.; Lobel, A.; Merle, T.; Royer, P.; Martayan, C.; David, M.; Hensberge, H.; Thienpont, E. The Belgian repository of fundamental atomic data and stellar spectra (BRASS). I. Cross-matching atomic databases of astrophysical interest. *Astron. Astrophys.* **2018**, *612*, A60. [CrossRef]