Assessment of the FAIRness of the Virtual Atomic and Molecular Data Centre following the Research Data Alliance evaluation framework

Carlo Maria Zwölf, Nicolas Moreau

Assessment of the FAIRness of the Virtual Atomic and Molecular Data Centre following the Research Data Alliance evaluation framework

Carlo Maria Zwölf[1] and Nicolas Moreau[1]

[1]LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne University, UPMC Univ Paris 06, 5 Place Janssen, Meudon, 92190 , France.

Contributing authors:
carlo-maria.zwolf@observatoiredeparis.psl.eu;
nicolas.moreau@observatoiredeparis.psl.eu;

Abstract

This version of the article has been accepted for publication, after peer review but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: https://doi.org/10.1140/epjd/s10053-023-00649-x. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms

In this paper we present the result of the analysis made on the Virtual Atomic and Molecular Data Centre infrastructure, following the FAIR Data Maturity Model framework defined by the Research Data Alliance. After recalling the technical architecture of the VAMDC e-infrastructure (subsection 1.1), we will introduce the RDA FAIR evaluation framework (section 2) and define the methodology we adopt to perform our analysis (subsection 2.1). In subsection 2.2 we will present the result of this analysis (the fine-grained granularity analysis from which subsection 2.2 is discussed in Appendix A). After having identified some lines of work aimed at improving the FAIRness of VAMDC (section 3), we will conclude with some ideas for further developments of this work (section 4).

Keywords: Atomic data, Molecular data, databases, VAMDC, FAIR principles, RDA

1 Introduction

The Virtual Atomic and Molecular Data Centre (VAMDC[1]) is a technical [1] and political [2] framework, built since 2009 through two European FP7 grants[2] for implementing and sustaining a worldwide digital research e-infrastructure, which has been successfully operated and maintained for more than 10 years by the VAMDC Consortium [3]. At the date of writing this paper, the VAMDC e-infrastructure interconnects in an interoperable way ∼41 heterogeneous Atomic and Molecular (A&M) databases that are mainly used for the interpretation of astronomical spectra and for modeling in media of many fields of astrophysics. Other application fields include atmospheric physics, plasmas, fusion, and radiation damage. VAMDC offers a common entry point to all the federated databases through the VAMDC-portal[3], providing a set of tools to retrieve and handle the data [4]. 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.

1.1 The technical architecture of the infrastructure

From a technical point of view, the VAMDC e-infrastructure relies on four pillar components. In the following list of items we briefly describe their functioning and in Table 1 we will give more details about the standards adopted/implemented to set-up those core components.

- **The set of federated data-nodes** - The "V" of VAMDC stands for "virtual" in the sense that the e-infrastructure does not contain data: an ad hoc generic wrapping software, the *node-software* [5], transforms an autonomous database into a VAMDC federated database. In VAMDC jargon a federated database is a *data-node*. Each data-node accepts queries submitted in a standard grammar[4] and, by implementing an interoperable data access protocol derived from the International Virtual Observatory Alliance (IVOA[5]) Table Access Protocol[6], provides output formatted into a common output format called XML Schema for Atoms, Molecules and Solids (XSAMS)[6].
- **The Registry**[7] is a central service where data-nodes metadata are registered using IVOA community standard [7]. Those metadata include the

[1] https://vamdc.org
[2] http://www.vamdc-project.vamdc.eu and http://www.sup-vamdc.vamdc.org
[3] https://portal.vamdc.org
[4] The VAMDC SQL Subset, see https://standards.vamdc.eu
[5] https://ivoa.net
[6] https://www.ivoa.net/documents/TAP/20190927/
[7] https://registry.vamdc.eu

resolvable identifiers of each node, the contact details for the scientific and technical maintainers, the versions of the implemented standards and protocols.

- **The Species database**[8] is a centralised chemical species repository, weekly updated by automated collation of data from all the VAMDC data-nodes. It contains a list of all the species in each VAMDC database. Each species is uniquely identified by a couple of international chemical identifiers InChI/InChIKey [8] (cf. Table 1, Table 2) and is characterized by its chemical names, formula, stoichiometric formula, mass number and charge.
- **The Query Store**[9] is a central repository for all the queries answered by all the data-nodes [9]. Each time a given data-node responds to an incoming query, it notifies this action to the Query Store. The latter assigns a machine actionable persistent resolvable identifier (PID) [10] to the query (optionally this identifier may be a DOI) which has just been notified. The Query Store stores the original query, together with the version of the data-node, the versions of the standards implemented by the node, the XSAMS data file produced while serving the query, the list of publications used to compile the retrieved dataset. It is worth noting that the Query Store implements international standards, defined by the Research Data Alliance (RDA[10]), facilitating the data-citation [11, 12].

Figure 1 shows how these pillar components interact together when a user submits a query to extract data from the VAMDC e-infrastructure.

2 VAMDC and the FAIR Principles

Science is an incremental process, where new discoveries are based on solid, proven and known results. Innovation is therefore connected to the amount of information shared within our communities. Those general ideas have been at the basis of the formulation of the FAIR principles [15]: research data should be Findable, Accessible, Interoperable, Reusable.

VAMDC experts have been involved for several years in international data-sharing organization [3] and have anchored, ahead of the FAIR principles formal definition time, those principles into the architecture of the VAMDC e-infrastructure: with the four pillars described in the subsection 1.1, data are easy to discover and to find, they are accessible, interoperable and easy to reuse. In a sense, VAMDC implemented the FAIR principles before their formal definition.

In recent years the Research Data Alliance (RDA) has produced a standard set of guidelines, the *FAIR Data Maturity Model* [16], aimed both at data producers and at e-infrastructure actors in order to evaluate the FAIRness (i.e. the maturity with respects to the FAIR principles) of their resources. We are going to apply the RDA guidelines to re-assess the FAIRness of the various

[8] https://species.vamdc.eu
[9] https://cite.vamdc.eu
[10] https://www.rd-alliance.org

Infrastructure element	Nature of data	Implemented standards	Adopted unique identifier
Nodes	Atomic and Molecular processes available in federated databases	XSAMS 12.07; VAMDC TAP*, VAMDC SQL Subset 2 Query Language*; VAMDC controlled dictionaries; IVOA Support Interfaces 1.0; IVOA VODataService 1.1	IUAPAC InChIKey for species; UUID for queries
Registry	Description of Nodes	VAMDC controlled dictionaries; IVOA Registry Interface 1.0; IVOA VODataService 1.1; IVOA Identifiers 2.0	IVOID for node identification
Species database	Description of Chemical species available from Nodes	XSAMS 12.07; JSON	IVOID for node identification; IUAPAC InChIKey for species
Query Store	Description of queries served by the Nodes	RDA Scalable Dynamic Data Citation; RDA/WDS Scholix; DataCite Metadata Schema 4.4	UUID and DOI for queries

Table 1 Overview of the standards implemented in the VAMDC infrastructure core components. The column *Adopted unique identifier* describes what kind of identifiers are used in each configuration: InChIKey [8] is the chemical identification mechanism defined by the International Union of Pure and Applied Chemistry (https://iupac.org); IVOID is the resource identification mechanism defined by the *Identifiers 2.0* standard from the IVOA. UUID and DOI are identification mechanisms defined by ISO-norms (cf. Table 2). For each standard listed in the column *Implemented Standards*, Table 2 provides further details.

* The standard VAMDC TAP is derived from the IVOA Table Access Protocol Standard (https://www.ivoa.net/documents/TAP).
* The standard VAMDC SQL Subset 2 Query Language is derived from the SQL standard (https://www.iso.org/standard/63555.html).

components of VAMDC: after introducing the methodology used to perform the evaluation (subsection 2.1), we will present the results of the evaluation (subsection 2.2).

Remark 1 The RDA Fair Data Maturity Model is not the only evaluation framework available. Different communities and projects have produced their own FAIR metrics/indicators. We may enumerate, non exhaustively, the FAIRs-FAIR Data Object Assessment Metrics [17], the FAIR Maturity Indicators from FAIRsharing [18] (https://fairsharing.github.io/FAIR-Evaluator-FrontEnd) and the FAIR Enough data maturity indicators (https://fair-enough.semanticscience.org/collections/fair-enough-data). As it is highlighted in [19], the RDA Fair Data Maturity Model is very complete and its indicators encompass most of the criteria from the

Table 2 Further information about the standards listed in the third column of Table 1: for each standard we recall the community perimeter (e.g. generic international standards vs. disciplinary standards) and the link to the document where the standard is defined.

Standard Name	**Community Perimeter**	**Link to standard definition**
XSAMS 12.07	VAMDC Internal (VI)	https://standards.vamdc.eu/dataModel/vamdcxsams/index.html#vamdcxsamslanguage-index
VAMDC controlled dictionaries	VI	https://standards.vamdc.eu/dictionary
VAMDC TAP	VI	https://standards.vamdc.eu/dataAccessProtocol
VAMDC SQL Subset 2 Query Language	VI	https://standards.vamdc.eu/queryLanguage
IVOA Registry Interface 1.0	Astronomy and Astrophysics (A&A)	https://www.ivoa.net/documents/RegistryInterface/20091104/
IVOA VODataService 1.1	A&A	https://www.ivoa.net/documents/VODataService/20101202/
IVOA Support Interfaces 1.0	A&A	https://www.ivoa.net/documents/VOSI/20110531
IVOA Identifiers 2.0	A&A	https://ivoa.net/documents/IVOAIdentifiers/20160523/
JSON	Generic International Standards (GIS)	https://www.json.org
RDA Scalable Dynamic Data Citation	GIS	http://dx.doi.org/10.15497/RDA00016
RDA/WDS Scholix	GIS	https://doi.org/10.5281/zenodo.1120265
UUID	GIS	https://www.iso.org/fr/standard/62795.html
DOI	GIS	https://www.iso.org/fr/standard/43506.html
IUAPAC InChIKey	Chemistry	https://www.inchi-trust.org
DataCite Metadata Schema 4.4	GIS	https://doi.org/10.14454/3w3z-sa82

other evaluation frameworks. This is the reason why we chose the RDA evaluation framework to assess the FAIRness of VAMDC.

2.1 Methodology and typographical conventions

The *FAIR Data Maturity Model* defines a set of 41 indicators, coupled to priorities and evaluation methods for each FAIR principle. Its authors pay great attention to the distinctions between data and metadata; while we believe that this distinction is not that important (in our opinion this depends on the vision of the data consumer: the metadata for some may be the data for the others, and conversely, cf. remark 2) we have to adopt this distinction to perform the FAIR evaluation. We arbitrarily define the following criteria:

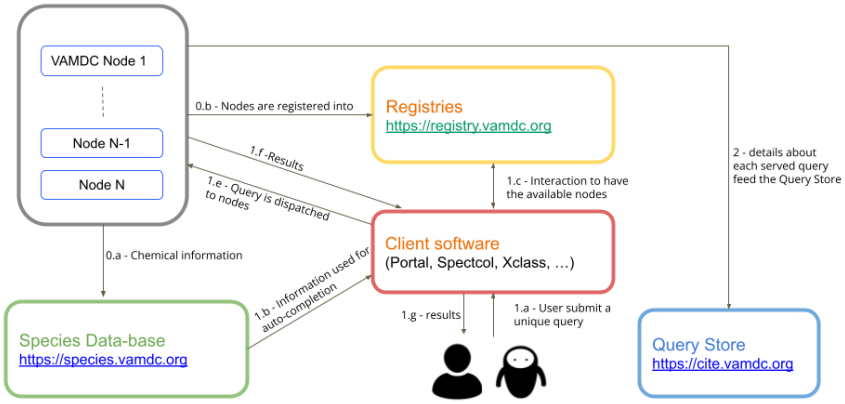

Fig. 1 Graphical representation of the information flow along a data extraction from the VAMDC e-infrastructure.
• Step 0 - indexing of chemical and technical information: chemical information from nodes (step 0.a) feeds the Species database. All the available nodes (step 0.b) are registered into the Registry. These processes are invisible to users and are part of the technical automated tasks that guarantee the functioning of the infrastructure.
• Step 1 - Query: when a user, who can be a human or a computer script, submit a query (step 1.a) through a VAMDC client software, chemical information from the Species database are used for auto-completion (step 1.b). Common clients are the Portal [4], Spectcol [13] or XClass [14]. The client receiving a query interacts with the Registry (step 1.c) to check what are the available nodes and then dispatches the query to all these nodes (step 1.e). Results from the nodes (step 1.f) are gathered by the client and served to the user (step 1.g).
• Step 2 - Query registration: the nodes having served the query notify to the Query Store the action they have just completed.

- **data**: we consider as data the output XSAMS file produced by a data-node while serving a given query;
- **metadata**: we consider as metadata all the other information contained in the Species database, in the Registry and in the Query Store.

Remark 2 One might object that the information contained into the Species database, due to its scientific nature, should be considered as data. Nevertheless we consider the content of the Species database as metadata since it is a synthetic view built from the chemical content of all the nodes : it is the consolidated and aggregated description of the chemical content of the nodes, and data describing data are commonly considered metadata.

In the subsection 2.2 we will provide an overview of the results of the evaluation derived from Appendix A where we assess in detail the FAIR Data Maturity Model indicators. We will use the following typographical convention:

- *Indicator Identifier*[11] *- Priority for the Indicator*

[11] we will use the same numbering convention introduced in [16]

$\mathcal{D}$ indicator short description;
$\mathcal{A}$ analysis for the indicator;
$\mathcal{L}$ the level of maturity for the indicator;
$\mathcal{I}$ when they are identified, our suggestions to improve the VAMDC FAIRness.

The Priority for the Indicator is part of the definition of the indicator and determines its importance. It may take one of these three values : Essential, Important, Useful. Following [16], the level of maturity for the indicator is an integer taking the following values: 0–not applicable; 1–not being considered yet; 2–under consideration or in planning phase: 3–in implementation phase; 4–fully implemented.

2.2 Results of evaluation

In this section we provide a synthetic view of the evaluation results, grouping the evaluation indicators by FAIR principle categories. The fine-grained analysis from which this section derives is detailed in Appendix A.

2.2.1 Evaluation for Findable principle

The evaluation for the **Findable** principle relies on 7 indicators (cf. Figure 2): all these seven indicators are essential. All except one indicator have a maturity level $\mathcal{L} = 4$. The only indicator having $\mathcal{L} = 3$ is *RDA-F4-01M* and is associated with metadata indexing and harvesting: metadata are well indexed withing the VAMDC ecosystem, but are not well indexed in generic services (e.g. Google).
With an average maturity level $\mathcal{L}_{avg}$ above 3.8, the analysis is very satisfactory for the Findable principle.

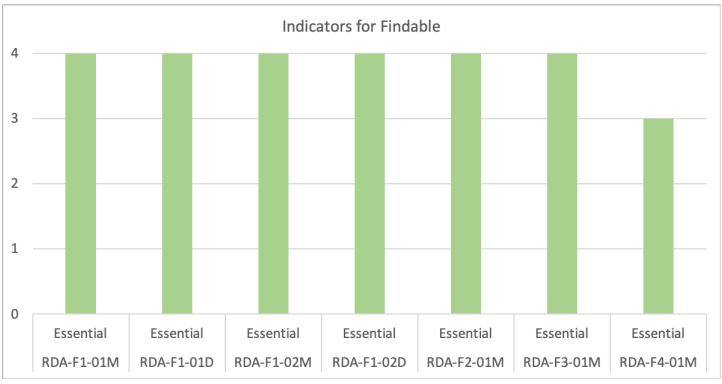

Fig. 2 Graphical representation of the analysis results for the **Findable** principle. 0 - not applicable; 1 - not being considered yet, 2 - under consideration or in planning phase; 3 - in implementation phase; 4 - fully implemented.

2.2.2 Evaluation for Accessible principle

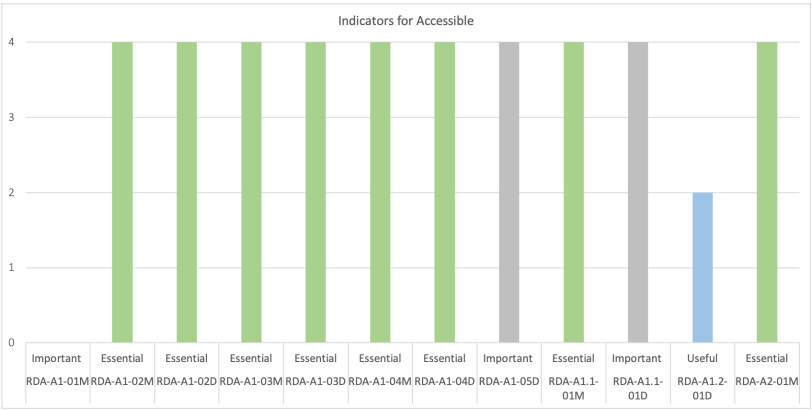

Fig. 3 Graphical representation of the analysis results for the **Accessible** principle. 0 - not applicable; 1 - not being considered yet, 2 - under consideration or in planning phase; 3 - in implementation phase; 4 - fully implemented. The color convention for the bars is: green for essential indicators, blue for useful and gray for important.

The evaluation for the **Accessible** principle relies on 12 indicators (cf. Figure 3): 8 are essential, 3 are important and 1 is useful. While one important indicator (*RDA-A1-01M*) does not apply for VAMDC, all the other essential and important indicators have a maturity level $\mathcal{L} = 4$. The useful indicator having $\mathcal{L} = 2$ is *RDA-A1.2-01D* and is associated with data accessibility through an access protocol that supports authentication and authorisation: the access protocol in VAMDC is completely free (no authentication required) but may be adapted to include an *authentication, authorization, accounting strategy* [20].
All the other indicators have a maturity level $\mathcal{L} = 4$. With an average maturity level $\mathcal{L}_{avg}$ above 3.8 ($\mathcal{L}_{avg} = 4$ for essential indicators, $\mathcal{L}_{avg} = 4$ for important indicators and $\mathcal{L} = 2$ for the unique useful indicator), the analysis is very satisfactory for the Accessible principle.

2.2.3 Evaluation for Interoperable principle

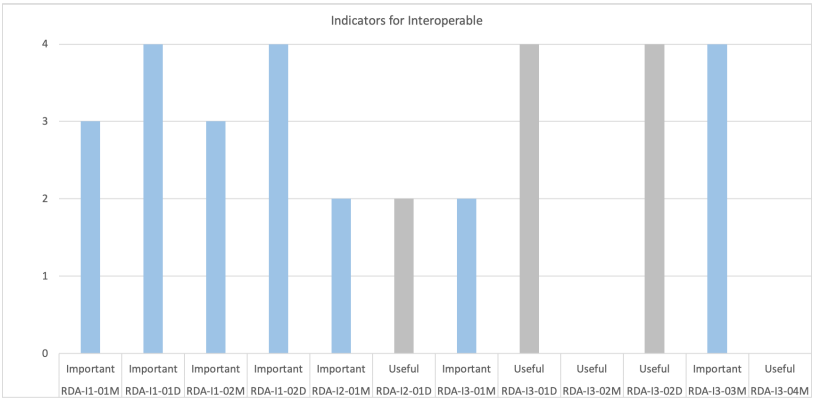

Fig. 4 Graphical representation of the analysis results for the **Interoperable** principle. 0 - not applicable; 1 - not being considered yet, 2 - under consideration or in planning phase; 3 - in implementation phase; 4 - fully implemented. The color convention for the bars is: green for essential indicators, blue for useful and gray for important.

The evaluation for the **Interoperable** principle relies on 12 indicators (cf. Figure 4): 7 are important and 5 are useful. Two indicators (*RDA-I3-02M* and *RDA-I3-04M*) do not apply for VAMDC.
The important indicators *RDA-I1-01M* and *RDA-I1-02M* (associated with the standard and machine actionable representation of metadata) have a maturity level $\mathcal{L} = 3$, because some service-endpoints on the Species database side and on the Query Store side are not standardised, nor machine actionable. The important indicator *RDA-I2-01M* and the useful indicator *RDA-I2-01D* (both associated with the usage of FAIR compliant vocabularies) have a maturity level $\mathcal{L} = 2$, because the dictionaries defined within the VAMDC infrastructure (cf. Table 2) are not fully FAIR compliant, since they lack machine actionability.

The important indicator *RDA-I3-01M* (about cross-references between metadata) has a maturity level $\mathcal{L} = 2$ because some references appearing on VAMDC user-interfaces are not machine actionable, but appear as simple text-strings.

All the other indicators have a maturity level $\mathcal{L} = 4$.

With an average maturity level $\mathcal{L}_{avg}$ around 3.2 ($\mathcal{L}_{avg} \approx 3.1$ for important indicators and $\mathcal{L}_{avg} \approx 3.3$ for useful indicators), the evaluation for the Interoperable principle is not as satisfactory as it is for the Findable and Accessible principles. However we may note that no essential indicator is involved and that the Interoperability may be greatly enhanced by systematically standardising the VAMDC service-endpoints and by adopting FAIR-compliant vocabularies.

2.2.4 Evaluation for Reusable principle

The evaluation for the **Reusable** principle relies on 10 indicators (cf. Figure 5): 5 are essential, 4 are important and 1 is useful.

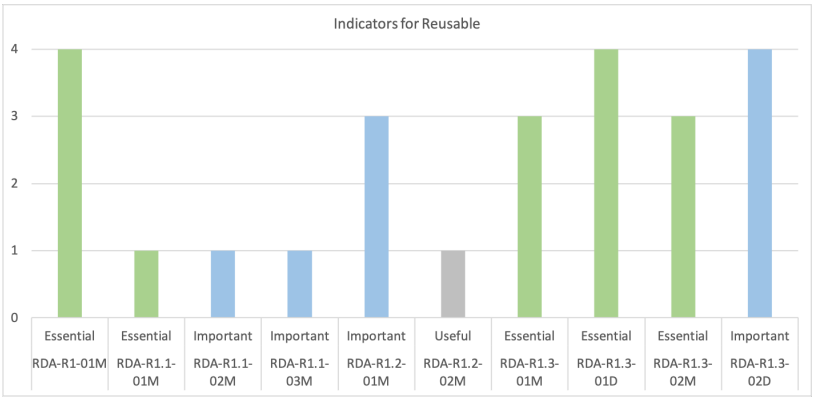

Fig. 5 Graphical representation of the analysis results for the **Reusable** principle. 0 - not applicable; 1 - not being considered yet, 2 - under consideration or in planning phase; 3 - in implementation phase; 4 - fully implemented. The color convention for the bars is: green for essential indicators, blue for useful and gray for important.

The essential indicator *RDA-R1.1-01M* and the two important indicators *RDA-R1.1-02M* and *RDA-R1.1-03M* (all these indicators are about licenses on data and metadata) have a maturity level $\mathcal{L} = 1$, because in VAMDC there is not a common policy about licenses and data come with no license information attached (cf. subsection 3.1 for a discussion about licenses issues in VAMDC).
The important indicator *RDA-R1.2-01M* and the useful indicator *RDA-R1.2-02M*, about provenance information and its expression using cross-community standards, have respectively a maturity level $\mathcal{L} = 3$ and $\mathcal{L} = 1$: the provenance information is not systematically timestamped (e.g. in the Species database) and the available provenance information does not follow any particular standard.
The two essential indicators *RDA-R1.3-01M* and *RDA-R1.3-02M* (about compliance of metadata with community and machine actionable standards) have a maturity level $\mathcal{L} = 3$ because some pieces of information returned by the Species database and the Query Store do not follow any metadata standard.
All the other indicators have a maturity level $\mathcal{L} = 4$.
With an average maturity level $\mathcal{L}_{avg}$ around 2.5 ($\mathcal{L}_{avg} = 3$ for essential indicators, $\mathcal{L}_{avg} \approx 2.2$ for important indicators and $\mathcal{L} = 1$ for the unique useful indicator), the evaluation for the Reusable principle is the worst compared to the previous one. Improvement can be easily made on the metadata compliance with community/machine actionable standards but are harder to realize on the license side (cf. subsection 3.1) and on the provenance side (cf. subsection 3.2).

3 Conclusion

The RDA FAIR Data Maturity Model, mainly designed for standalone data-sets or repositories, perfectly scales in the case of a distributed and complex e-infrastructure as VAMDC, with different level of data and metadata (cf. subsection 1.1). Although being generic (all the indicators are technology and implementation neutral), the RDA framework allows a very precise and specific analysis and leads to concrete ways of improvement.
The FAIRness level of VAMDC infrastructure is satisfactory for experts aware of

the infrastructure conventions and standards, but may be significantly increased for newcomers or for trans-disciplinary activities. The lines of improvement identified through the subsection 2.2 (and Appendix A) may be summarized as follow:

- A single entry point should be provided to get all the information about a given node. Today this information is fragmented between the Species database and the Registry;
- As discussed while evaluating the indicator RDA-I3-01M, the cross-references of metadata between the different pillar components (cf. subsection 1.1) should be in the form of persistent resolvable machine actionable identifiers;
- Register VAMDC standards and formats into ad hoc registries of types (e.g. FAIRsharing ones) and assign persistent resolvable machine actionable identifiers to each standard in order to easily refer to it;
- Systematically use FAIR compliant dictionaries to express knowledge;

The main obstacles that explain the unsatisfactory evaluation for the Reusable principle (cf. subsubsection 2.2.4), are related to the absence of a license-policy and the non-adoption of a standard representation for provenance information. We are going to discuss these two aspects respectively in subsection 3.1 and subsection 3.2.

3.1 Discussion about licenses

The absence of licenses is a legal but not a technical obstacle to the reuse of data. Legal considerations are very complex in the case of VAMDC, because of the international character of the Consortium.

Let us first focus on the data aspects: VAMDC data is stored in standalone databases (the federated data-Nodes, cf. subsection 1.1) which are located in different countries. The database right (which is a form of property right applied to compilation of data in a database) is recognised only in a small number of jurisdiction (e.g. in Europe but not in the United States of America) and, even in cases where this right is recognized, it is interpreted differently in different states (cf. for example https://en.wikipedia.org/wiki/Database_right, accessed on the 23rd of January 2023). Without a common legal framework for discussions, it is very difficult for the data providers as well as the VAMDC maintainers to agree on a suitable data-licence.

Let us now consider the metadata aspects: the VAMDC metadata is stored in the Registry, the Species database and the Query Store, which are central services. These services are hosted and subject to the legislation of the legal representative on the VAMDC Consortium. The role of legal representative is not open-ended and several legal representatives may succeed one another over the years, moving from one jurisdiction to another. As in the case of data, it is very difficult to converge to a Consortium agreement on metadata licenses: this must be based on the specific jurisdiction of the legal representative's state but must be approved by all the international partners. This is even harder if we consider that the legal framework may change with the change of the legal representative.

It is important to note that, for some data producers, these discussions are merely legal quibbles that have little to do with science and the actual reusability of their

data. International alliances, such as RDA and CODATA[12] and their *Legal Interoperability Interest Group* [13], can raise awareness of the importance of legal issues and provide suitable solutions for international data collaborations.

3.2 Discussion about provenance

Provenance information is crucial in any scientific process: if a researcher wants to reuse some data, he/she needs to understand how the data were produced, what their domains of validity and application are. In VAMDC the provenance information for data and metadata is well documented and made available to users (cf. section 2.2.4), except for the lack of some time-stamped metadata information that we reported during the review of indicator *RDA-R1.2-01M*.
As we highlighted while evaluating the indicator *RDA-R1.2-02M*, VAMDC provenance information does not follow any particular format or standard (as for example W3C[14] or IVOA-prov[15]). One might wonder whether it would be useful for VAMDC to adopt these provenance standards. On the one hand, we must remember that the indicator *RDA-R1.2-02M* is just considered "useful". On the other hand, the provenance standards are very technical [21], and data producers should make a considerable effort to implement them. Moreover, the high degree of technicality would make the provenance information difficult to understand for the end users[16]: ad hoc tools should be provided to help users in decoding the provenance information from the standard to a more human friendly representation. The provenance standards are designed and extremely useful in the context of a fully automated data processing, such as scientific/computational workflows [22], but workflow engines are not yet part of the practices of most VAMDC users.
Considering the elements we have just exposed, we believe that it would be premature to implement provenance standards on the VAMDC infrastructure and that the information currently provided fulfils users' needs.

4 Final remarks and further works

A possible continuation of this work would be to implement the improvement lines highlighted in section 3. This will depend on the priorities of the VAMDC Consortium, which are based on an analysis of the needs of VAMDC community of data providers and users, and on the available development manpower and funding.

Automatic tools to evaluate the FAIR maturity level of data-repositories have been developed recently [23, 24]. Following an approach similar to the one described in [25] we may test some automatic tool on VAMDC: it could be interesting to discover how those automatic tools, mainly built for standalone data repositories, will work in the case of VAMDC, which is a network of distributed dynamic databases orchestrated via a complex infrastructure with different levels of identifiers and metadata (cf. subsection 1.1). It will also be interesting to compare the results from the automatic tools with the analysis presented in this work.

[12] https://codata.org
[13] https://www.rd-alliance.org/groups/rdacodata-legal-interoperability-ig.html
[14] https://www.w3.org/TR/prov-overview/
[15] https://www.ivoa.net/documents/ProvenanceDM/
[16] cf. https://github.com/mservillat/voprov/blob/master/tutorials/voprov_tutorial.ipynb for an example of how provenance information formatted using IVOA-prov appears to end users

Appendix A Details of the evaluation

- *RDA-F1-01M - Essential*

  $\mathcal{D}$ **Metadata are identified by a persistent identifier**;
  $\mathcal{A}$ VAMDC federates heterogeneous databases, each one of those has its own PID strategy and procedure for metadata. Moreover these databases are dynamic. To address the issue of persistently identifying metadata associated with VAMDC data-extractions we implemented the RDA Dynamic Data Citation recommendation [9]: the Query Store achieves the persistent identification of both metadata and data extracted from the VAMDC infrastructure;
  $\mathcal{L} = 4$

- *RDA-F1-01D - Essential*

  $\mathcal{D}$ **Data is identified by a persistent identifier**;
  $\mathcal{A}$ The analysis is the same as in the case of the indicator *RDA-F1-01M*;
  $\mathcal{L} = 4$

- *RDA-F1-02M - Essential*

  $\mathcal{D}$ **Metadata are identified by a globally unique identifier**;
  $\mathcal{A}$ The Query Store assigns to any extracted digital object (i.e. the data together with its relevant metadata) a UUID[17] and, upon user request, a DOI. These identifiers are unique by construction;
  $\mathcal{L} = 4$

- *RDA-F1-02D - Essential*

  $\mathcal{D}$ **Data is identified by a globally unique identifier**;
  $\mathcal{A}$ The analysis is the same as in the case of *RDA-F1-02M*;
  $\mathcal{L} = 4$

- *RDA-F2-01M - Essential*

  $\mathcal{D}$ **Rich metadata are provided to allow discovery**;
  $\mathcal{A}$ As we discussed in subsection 2.1, VAMDC provides rich metadata to describe its data. We identified three different sets of metadata within the infrastructure, each set being associated to a particular infrastructure component:

  1. The metadata associated with the served queries characterize the queries (e.g. the query-execution time, the query text-string, the version of nodes having generated the data, the version of the standard

[17] https://www.iso.org/fr/standard/53416.html

used to format the output data, etc...). These metadata are stored on the Query Store;
2. The metadata associated with the available resources (e.g. list of supported requestable, returnable and restrictables[18], node-capabilities, supported access protocols, names of node maintainers, contacts, etc...). These metadata constitute the data stored into the Registry;
3. The metadata about the species and processes available for each node. These metadata constitute the data stored into the Species Database;

These three sets of metadata help users in discovering which database may contain interesting information.

$\mathcal{L}$ = 4

$\mathcal{I}$ A better integration between the Species database information and those from the Registry is encouraged. Users should have a single access point giving all the information on a given node, without having to use two different services. Today the information is fragmented.

- *RDA-F3-01M - Essential*

$\mathcal{D}$ **Metadata include the identifier for the data**;

$\mathcal{A}$ 1. In the case of the Query Store, the PID associated to the query resolves to a Digital Object containing both the metadata and the underlying data. These are accessible for both human and machines;
2. In the case of the Registry, the underlying data associated with the metadata are the resources (e.g. nodes and processors). These are clearly identified by their IVOID (cf. Table 1) and their service endpoint;
3. In the case of the Species database, the underlying data associated with the metadata are the nodes and the processes supported by the nodes. In the web-human interface, these are clearly identified (by a direct link to the autonomous version of the database, not the VAMDC IVOID or TAP endpoint) whereas in the Excel machine oriented interface[19], the IVOID of resources are provided. The link with the processes is a simple text string (only human readable information in the web interface) and is missing in the machine actionable interface;

$\mathcal{L}$ = 4

$\mathcal{I}$ In the case of the Species database, improve the machine actionability of the information, by putting for example links to the VAMDC IVOID of ressources in the human web-interface. Also pointers to processes should be machine actionable.

- *RDA-F4-01M - Essential*

[18] Requestable, returnable and restrictables (https://standards.vamdc.eu/dictionary/index.html) are three controlled vocabularies containing specific terms that clients may use to interact with the data-nodes.

[19] https://species.vamdc.org/toxls

$\mathcal{D}$ **Metadata are offered in such a way that they can be harvested and indexed**;

$\mathcal{A}$ The above described three sets of metadata are offered in a way which may be indexed. VAMDC itself indexes those metadata and provides discipline-oriented (atomic and molecular data community) services and portal. However it does not seem that the metadata from VAMDC are harvested and/or indexed outside VAMDC (e.g. Google or other indexing services). From the technical point of view, there an no particular barriers to achieve a better integration between indexing services and VAMDC, following an approach similar to the one described in [26];

$\mathcal{L} = 3$

$\mathcal{I}$ VAMDC experts may define strategies to improve metadata harvesting and indexing from outside their data services. Two different strategies may be sketched: prepare metadata in formats that may be directly ingested by indexing services [27], or register VAMDC endpoints and standards in public catalogues of services (e.g. EOSC Portal[20], FAIRsharing [28], etc...) so that harvesting services may discover and interact with VAMDC resources.

- *RDA-A1-01M - Important*

$\mathcal{D}$ **Metadata contain information to enable the user to get access to the data**;

$\mathcal{A}$ The data in VAMDC are completely free and open. In the case of the Query Store, the digital object may be accessed directly by resolving a given PID. In all the other cases data are accessible following classic https links. Since the access to the data is straightforward (in the sense that no authentication or particular procedure is required to use the services) VAMDC does not require specific metadata to enable users to get access to the data.

$\mathcal{L} = 0$

$\mathcal{I}$ It could be useful for users to systematically provide information about the conditions of usage for VAMDC services, personal data & privacy policy and licenses on data. Those conditions are already defined for the Portal and for the Query Store. These should be extended to the Species database and to the Registry cf. indicator *RDA-R1.1-01M*.

- *RDA-A1-02M - Essential*

$\mathcal{D}$ **Metadata can be accessed manually (i.e. with human intervention)**;

$\mathcal{A}$ The core central services of VAMDC (cf. subsection 1.1) have a web interface designed for humans. All the metadata are accessible through those interfaces;

$\mathcal{L} = 4$

[20] https://eosc-portal.eu

- *RDA-A1-02D - Essential*

  $\mathcal{D}$ **Data can be accessed manually (i.e. with human intervention)**;
  $\mathcal{A}$ The VAMDC portal [4] have a web interface designed for humans. All the data are accessible through those interfaces;
  $\mathcal{L} = 4$

- *RDA-A1-03M - Essential*

  $\mathcal{D}$ **Metadata identifier resolves to a metadata record**;
  $\mathcal{A}$ 1. In the case of the Query Store, the PID resolves directly to a digital object where the metadata and the underlying data are packed together. Resolving a Query Store PID gives access both to query metadata and query produced data;
  2. In the case of the registries, a resolution service is implemented which, starting from the resources IVOID (cf. Table 1) resolves to the underlying Registry node-resource record;
  3. In the case of the Species Database, a resolution service is implemented which, starting from a given species identifier resolves to the underlying meta-data;
  $\mathcal{L} = 4$
  $\mathcal{I}$ The endpoints of machine oriented resolution services for the Species database and the Registry should be advertised and indexed in some ad hoc service, so everybody may discover these endpoints.

- *RDA-A1-03D - Essential*

  $\mathcal{D}$ **Data identifier resolves to a digital object**;
  $\mathcal{A}$ In the case of the Query Store, the PID resolves directly to a digital object where the metadata and underlying data are packed together. Resolving a Query Store PID gives access both to query metadata and to the extracted data;
  $\mathcal{L} = 4$

- *RDA-A1-04M - Essential*

  $\mathcal{D}$ **Metadata are accessed through standardised protocol**;
  $\mathcal{A}$ All the communications between the VAMDC web services are built over the http protocol (and its secured version https), which is free and standard. All the data and metadata are accessed using http. If we go deeply into details, the messages encapsulated into this http envelope also follow standard and open protocols : the Registry relies over the IVOA recommendation[21], the data-nodes implements the TAP protocol

[21] https://www.ivoa.net/documents/RegistryInterface/20091104/

for data extraction[22]and accepts queries formatted into a standard query language[23];
$\mathcal{L}$ = 4

- *RDA-A1-04D - Essential*

  $\mathcal{D}$ **Data are accessible through standardised protocol**;
  $\mathcal{A}$ The analysis is the same as in the case of *RDA-A1-04M*;
  $\mathcal{L}$ = 4

- *RDA-A1-05D - Important*

  $\mathcal{D}$ **Data can be accessed automatically (i.e. by a computer program)**;
  $\mathcal{A}$ The architecture of VAMDC infrastructure relies on web services designed for machine-to-machine interaction. The web interfaces and tools designed for humans are an overlay to those services. Data extraction and handling is fully machine actionable;
  $\mathcal{L}$ = 4

- *RDA-A1.1-01M - Essential*

  $\mathcal{D}$ **Metadata are accessible through a free access protocol**;
  $\mathcal{A}$ The analysis is the same as in the case of *RDA-A1-04M*;
  $\mathcal{L}$ = 4

- *RDA-A1.1-01D - Important*

  $\mathcal{D}$ **Data is accessible through a free access protocol**;
  $\mathcal{A}$ The analysis is the same as in the case of *RDA-A1-04M*;
  $\mathcal{L}$ = 4

- *RDA-A1.2-01D - Useful*

  $\mathcal{D}$ **Data are accessible through an access protocol that supports authentication and authorisation**;
  $\mathcal{A}$ As previously mentioned, all the VAMDC data and services are open and free. However, as discussed in [20], the protocols may support authentication and authorisation;
  $\mathcal{L}$ = 2

- *RDA-A2-01M - Essential*

  $\mathcal{D}$ **Metadata are guaranteed to remain available after data is no longer available**;

[22] https://standards.vamdc.eu/dataAccessProtocol/index.html
[23] https://standards.vamdc.eu/queryLanguage/index.html

$\mathcal{A}$ The Species Database and the Registry are services describing the real-time state of the infrastructure. The legacy service for VAMDC data is the Query Store. Its metadata are kept permanently, even after the data deletion;

$\mathcal{L} = 4$

- *RDA-I1-01M - Important*

$\mathcal{D}$ **Metadata use knowledge representation expressed in standardised format**;

$\mathcal{A}$ 1. The DOI associated with a query in the Query Store resolves to a Digital Object whose metadata follow the DataCite metadata format[24]. The metadata, obtained while resolving the PID on the Query Store side, are formatted in JSON text, but do not follow any specific standard;
2. The metadata in the Registry follow the IVOA standards already cited and the information provided is built over standard dictionaries[25];
3. The metadata in the Species database do not use specific standards to express knowledge;

$\mathcal{L} = 3$

$\mathcal{I}$ 1. The format of the metadata returned by the Query Store while resolving a PID should be documented (both in human readable and machine actionable way) and published both through the VAMDC standards and in metadata registries, such as FAIRsharing[26][28];
2. In the Registry, when a specific term is used (for example a given value for a restrictable or returnable) there should be a pointer, under the form of a PID, resolving to the definition of the term itself. A less desirable solution should be to point to the VAMDC-dictionary repository from the Registry. Also, the version of standards should contain a PID resolving to the standards definition. Up to now, the links between the access protocol, the dictionaries and the other components are explained in human readable documentations but not explicitly expressed in machine interfaces;
3. A dictionary specific to the knowledge representation of the Species database should be introduced and published (both in human readable and machine actionable way), both on the VAMDC standard page and in metadata registries such as FAIRsharing.

- *RDA-I1-01D - Important*

$\mathcal{D}$ **Data use knowledge representation expressed in standardised format**;

[24] https://doi.org/10.14454/3w3z-sa82
[25] https://standards.vamdc.eu/dictionary
[26] https://fairsharing.org/search?fairsharingRegistry=Standard

$\mathcal{A}$ The data extracted from VAMDC are formatted using the community specific XSAMS standard. The XSAMS standard is documented in a human readable form and in a machine-actionable form via an XSD schema;
$\mathcal{L} = 4$
$\mathcal{I}$ Three aspects may improve the interoperability of the XSAMS standard:

1. Register the XSAMS standard as a standard format into registries of types such as FAIRsharing;
2. The VAMDC format contains enumerations of terms which are defined neither in the scheme, nor in an external dictionary: for example *CategoryType* in *Methods* object, *dataDescription* values in *DataSets* object;
3. For codes of processes, PID resolving to the definition of the codes should be introduced. In particular for IAEA DCN codes[27], the relevant metadata associated to each code is resolvable at https://amdis.iaea.org/databases/processes/⟨CODE⟩ as a JSON object (e.g. https://amdis.iaea.org/databases/processes/HCX/).

- *RDA-I1-02M - Important*

$\mathcal{D}$ **Metadata use machine-understandable knowledge representation**;
$\mathcal{A}$ 1. The DOI associated with a query in the Query Store resolves to a Digital Object whose metadata follow the DataCite metadata format, which is machine actionable. The metadata obtained while resolving the PID on the Query Store side are formatted in JSON text, but do not follow any specific standard. Metadata are machine-readable but not machine-understandable;
2. Metadata in registry are built to be machine-understandable;
3. Metadata in the Species database are formatted in JSON text, but do not follow any specific standard. Metadata are machine-readable but not machine-understandable;

$\mathcal{L} = 3$
$\mathcal{I}$ Improve the machine-understandability of the information from the Query Store and from the Species database.

- *RDA-I1-02D - Important*

$\mathcal{D}$ **Data use machine-understandable knowledge representation**;
$\mathcal{A}$ The XSAMS standard (cf. subsection 1.1) is, as its name indicates, an XML schema which is designed to be machine understandable. VAMDC provides several tools and converter to automatically interact with XSAMS data;
$\mathcal{L} = 4$

[27] These are codes of processes historically defined by the Data Centre Network initiative at the IAEA (the legacy web site is https://www-amdis.iaea.org/DCN/). Today the activity is led by the AMD unit (https://amdis.iaea.org/databases/processes/)

- *RDA-I2-01M - Important*

  $\mathcal{D}$ **Metadata use FAIR-compliant vocabularies**;
  $\mathcal{A}$ 1. The DOIs related to query store resolve to Digital Objects which use DataCite metadata (cf. Table 1 and Table 2), and as a consequence dictionaries that are compliant with the Dublin Core extended dictionary[28]. The PIDs of the Query Store use no FAIR dictionaries;
  2. The registries use elements from VAMDC dictionaries to define capacities of nodes (e.g. Restrictables and returnables). However only the label of the term is used. There is no pointer to a PID resolving to the term definition. Moreover the VAMDC dictionaries are not fully FAIR since they are not designed for machine but only for human fruition (e.g. There is just a web site and no API to access the terms and their definition);
  3. The species DB does not use dictionaries to define the nature of metadata;

  $\mathcal{L} = 2$
  $\mathcal{I}$ Make the VAMDC dictionaries more FAIR by improving machine-actionability and term identification. Introduce dictionaries for the Species database.

- *RDA-I2-01D - Useful*

  $\mathcal{D}$ **Data use FAIR-compliant vocabularies**;
  $\mathcal{A}$ The VAMDC format contains enumeration of terms which are defined neither in the scheme, nor in an external dictionary, e.g. *CategoryType* in *Methods* object or *dataDescription* values in *DataSet* object. The codes for processes (including IAEA DCN codes) are simple text strings;
  $\mathcal{L} = 2$
  $\mathcal{I}$ Improvements identified at item $\mathcal{I}$-2 and $\mathcal{I}$-3 of indicator *RDA-I1-01D* apply also here.

- *RDA-I3-01M - Important*

  $\mathcal{D}$ **Metadata includes references to other metadata**;
  $\mathcal{A}$ 1. The Query Store provides information about the node generating the query and the version of the standard. These references are only human readable text-string;
  2. In the Registry, the records of data-nodes contain some references to other metadata (services endpoints, curators, versions of standard) in a machine understandable way, but not necessary machine-actionable: those references cannot be resolved, as for example in the registry entry for Basecol[29] [29] one can find these two fields:

[28] A standard dictionary built for the semantic web, cf. https://www.iso.org/standard/71339.html and https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[29] https://registry.vamdc.eu/registry-12.07/main/viewResourceEntry.jsp?IVORN=ivo://vamdc/basecol2015/vamdc-tap

```
<versionOfStandards>
    12.07
</versionOfStandards>
```

and

```
<versionOfSoftware>
   Java VAMDC - TAP node implementation
    v.12.07r3 - SNAPSHOT
</versionOfSoftware>
```

We see that we have no direct link to the definition of the cited standard, nor to the repository for the cited software. The two text-strings enclosed by the XML tags may be understood only by expert users;

3. The Species database contains references to the nodes providing a given species. By now this points directly to the the historical web sites of databases as they existed (and still exist) before the building of VAMDC and not to the VAMDC data-node versions (i.e. to the corresponding data-node resources, as defined in the Registry);

$\mathcal{L} = 2$

$\mathcal{I}$ 1. The Query Store should provide references pointing to the node record in the registries and to definition of standards;
2. In the Registry, the metadata should point to definition of standards or to the software repositories in a machine actionable way;
3. The Species database should have a link to the Registry record of each node.

- *RDA-I3-01D - Useful*

$\mathcal{D}$ **Data include references to other data**;

$\mathcal{A}$ In the XSAMS schema, the *Source* object may be used to reference other data and/or papers playing the role of source for the current data;

$\mathcal{L} = 4$

$\mathcal{I}$ By now it does not exist a mechanism to reference connected data (e.g. similar data, previous version of the same data, etc..) directly inside XSAMS. VAMDC should investigate if such a need exists in its community.

- *RDA-I3-02M - Useful*

$\mathcal{D}$ **Metadata include references to other data**;

$\mathcal{A}$ The DOIs assigned by the Query Store are versioned. The filiation of data with previous version are a typical example of "metadata include references to other data". In all the other cases this indicator is not applicable;

$\mathcal{L} = 0$

- *RDA-I3-02D - Useful*

$\mathcal{D}$ **Data include qualified references to other data**;
$\mathcal{A}$ The analysis is the same as for indicator *RDA-I3-01D*. The references are qualified in a machine readable way (but not necessarily actionable since the DOI is not mandatory in the XSAMS *Source* object) as "sources for the data";
$\mathcal{L} = 4$

- *RDA-I3-03M - Important*

$\mathcal{D}$ **Metadata include qualified references to other metadata**;
$\mathcal{A}$ The references described while evaluating *RDA-I3-01M* are all qualified in a human oriented way and in a machine readable way;
$\mathcal{L} = 4$
$\mathcal{I}$ Metadata references to other metadata should be machine actionable, cf. the analysis of *RDA-R1-01M*.

- *RDA-I3-04M - Important*

$\mathcal{D}$ **Metadata include qualified references to other data**;
$\mathcal{A}$ The analysis is the same as for indicator *RDA-I3-02M*;
$\mathcal{L} = 0$

- *RDA-R1-01M - Essential*

$\mathcal{D}$ **Plurality of accurate and relevant attributes are provided to allow reuse**;
$\mathcal{A}$ As explained while evaluating the previous indicators, VAMDC uses a wide number of metadata to describe its data. The quality of those metadata is high;
$\mathcal{L} = 4$
$\mathcal{I}$ Use standards and norm for all metadata to get them machine actionable.

- *RDA-R1.1-01M - Essential*

$\mathcal{D}$ **Metadata include information about the licence under which the data can be reused**;
$\mathcal{A}$ There is no metadata describing the license on data. Data in VAMDC are free, but the consortium never discussed about a license to apply on VAMDC extracted data. As of today, there is no metadata field defined to specify the adopted license;
$\mathcal{L} = 1$
$\mathcal{I}$ Define a license for all the data/metadata in VAMDC (cf. subsection 3.1).

- *RDA-R1.1-02M - Important*

$\mathcal{D}$ **Metadata refer to a standard reuse licence**;

$\mathcal{A}$ Since there are no license (and no metadata field to provide license information), we cannot point to a standard reuse license;
$\mathcal{L} = 1$
$\mathcal{I}$ The adopted license for data/metadata should be a widely adopted and accepted one (e.g. one from the Creative Commons ecosystem[30]).

- *RDA-R1.1-03M - Important*

$\mathcal{D}$ **Metadata refers to a machine-understandable reuse licence**;
$\mathcal{A}$ The analysis is the same as for indicator *RDA-R1.1-02M*;
$\mathcal{L} = 1$
$\mathcal{I}$ The adopted license should be machine actionable.

- *RDA-R1.2-01M - Important*

$\mathcal{D}$ **Metadata includes provenance information according to community-specific standards**;
$\mathcal{A}$ We recall that provenance metadata in the XSAMS is well defined using the *Source* element;

1. Provenance of data in the Query Store is clearly stated, since all the metadata of a landing page are about provenance. This information does not follow any particular standard but are human readable;
2. Provenance of data in the registries is well defined. The information follows IVOA standard;
3. Provenance in the Species Database is well defined (we know from which node the information comes). However we miss the timestamp information about the databases that provide the species information.

$\mathcal{L} = 3$
$\mathcal{I}$ Adopt standards for provenance in the Query Store and in the Species database.

- *RDA-R1.2-02M - Useful*

$\mathcal{D}$ **Metadata include provenance information according to a cross-community language**;
$\mathcal{A}$ Cross community standard from provenance (W3C[31] or IVOA-prov[32]) are not used (cf. subsection 3.2 for a discussion);
$\mathcal{L} = 1$

- *RDA-R1.3-01M - Essential*

$\mathcal{D}$ **Metadata comply with a community standard**;

[30] https://creativecommons.org
[31] https://www.w3.org/TR/prov-overview/
[32] https://www.ivoa.net/documents/ProvenanceDM/

$\mathcal{A}$ 1. The DOIs assigned by the Query Store use metadata compliant with the DataCite metadata standard;
2. The metadata constituting the registries follows the IVOA standard for registries;
3. In the Species database, we use Inchi and InchiKey chemical standards for species identifications. The other information do not follow any particular standard;

$\mathcal{L} = 3$

- *RDA-R1.3-01D - Essential*

$\mathcal{D}$ **Data comply with a community standard**;
$\mathcal{A}$ Data extracted from VAMDC are formatted using the XSAMS standard;
$\mathcal{L} = 4$

- *RDA-R1.3-02M - Essential*

$\mathcal{D}$ **Metadata are expressed in compliance with a machine-understandable community standard**;
$\mathcal{A}$ 1. The DOIs assigned by the Query Store use metadata compliant with the DataCite standard, which are machine actionable;
2. Metadata constituting the Registry follow the IVOA standard for registries, which are machine understandable and actionable;
3. The Species database metadata do not follow any particular standard;

$\mathcal{L} = 3$
$\mathcal{I}$ Use machine actionable standards for the information returned by the Query Store and by the Species database.

- *RDA-R1.3-02D - Important*

$\mathcal{D}$ **Data are expressed in compliance with a machine-understandable community standard**;
$\mathcal{A}$ Data extracted from VAMDC are formatted using the XSAMS standard, which is fully machine understandable;
$\mathcal{L} = 4$
$\mathcal{I}$ Register XSAMS into registries of types [30], such as FAIRsharing[28].

Acknowledgments. The authors thank Prof. M.L. Dubernet, previous chair of VAMDC, for suggesting the investigation of FAIR principles in VAMDC. They also thank the authors of [16] for their ready availability to answer questions related to the interpretation of indicators and the anonymous reviewers for their advice, which helped to improve the manuscript considerably.

Authors Contribution Statement. The first author wrote all the manuscript. The second author helped in filling the content of Table 1 and Table 2.

Data Availability Statement. Data sharing not applicable to this article as no datasets were generated during the current study. The data and metadata analysed during the current study are publicly available through the VAMDC infrastructure, following the HTML links systematically cited in the text.

References

[1] Dubernet, M.L., Boudon, V., Culhane, J.L., Dimitrijevic, M.S., Fazliev, A.Z., Joblin, C., Kupka, F., Leto, G., Le Sidaner, P., Loboda, P.A., Mason, H.E., Mason, N.J., Mendoza, C., Mulas, G., Millar, T.J., Nuñez, L.A., Perevalov, V.I., Piskunov, N., Ralchenko, Y., Rixon, G., Rothman, L.S., Roueff, E., Ryabchikova, T.A., Ryabtsev, A., Sahal-Bréchot, S., Schmitt, B., Schlemmer, S., Tennyson, J., Tyuterev, V.G., Walton, N.A., Wakelam, V., Zeippen, C.J.: Virtual atomic and molecular data centre. Jqsrt **111**, 2151–2159 (2010). https://doi.org/10.1016/j.jqsrt.2010.05.004

[2] Dubernet, M.L., Antony, B.K., Ba, Y.A., Babikov, Y.L., Bartschat, K., Boudon, V., Braams, B.J., Chung, H.-K., Daniel, F., Delahaye, F., Del Zanna, G., de Urquijo, J., Dimitrijević, M.S., Domaracka, A., Doronin, M., Drouin, B.J., Endres, C.P., Fazliev, A.Z., Gagarin, S.V., Gordon, I.E., Gratier, P., Heiter, U., Hill, C., Jevremović, D., Joblin, C., Kasprzak, A., Krishnakumar, E., Leto, G., Loboda, P.A., Louge, T., Maclot, S., Marinković, B.P., Markwick, A., Marquart, T., Mason, H.E., Mason, N.J., Mendoza, C., Mihajlov, A.A., Millar, T.J., Moreau, N., Mulas, G., Pakhomov, Y., Palmeri, P., Pancheshnyi, S., Perevalov, V.I., Piskunov, N., Postler, J., Quinet, P., Quintas-Sánchez, E., Ralchenko, Y., Rhee, Y.-J., Rixon, G., Rothman, L.S., Roueff, E., Ryabchikova, T., Sahal-Bréchot, S., Scheier, P., Schlemmer, S., Schmitt, B., Stempels, E., Tashkun, S., Tennyson, J., Tyuterev, V.G., Vujčić, V., Wakelam, V., Walton, N.A., Zatsarinny, O., Zeippen, C.J., Zwölf, C.M.: The virtual atomic and molecular data centre (VAMDC) consortium. Journal of Physics B Atomic Molecular Physics **49**(7), 074003 (2016). https://doi.org/10.1088/0953-4075/49/7/074003

[3] Albert, D., Antony, B.K., Ba, Y.A., Babikov, Y.L., Bollard, P., Boudon, V., Delahaye, F., Del Zanna, G., Dimitrijević, M.S., Drouin, B.J., Dubernet, M.-L., Duensing, F., Emoto, M., Endres, C.P., Fazliev, A.Z., Glorian, J.-M., Gordon, I.E., Gratier, P., Hill, C., Jevremović, D., Joblin, C., Kwon, D.-H., Kochanov, R.V., Krishnakumar, E., Leto, G., Loboda, P.A., Lukashevskaya, A.A., Lyulin, O.M., Marinković, B.P., Markwick, A., Marquart, T., Mason, N.J., Mendoza, C., Millar, T.J., Moreau, N., Morozov,

S.V., Möller, T., Müller, H.S.P., Mulas, G., Murakami, I., Pakhomov, Y., Palmeri, P., Penguen, J., Perevalov, V.I., Piskunov, N., Postler, J., Privezentsev, A.I., Quinet, P., Ralchenko, Y., Rhee, Y.-J., Richard, C., Rixon, G., Rothman, L.S., Roueff, E., Ryabchikova, T., Sahal-Bréchot, S., Scheier, P., Schilke, P., Schlemmer, S., Smith, K.W., Schmitt, B., Skobelev, I.Y., Srecković, V.A., Stempels, E., Tashkun, S.A., Tennyson, J., Tyuterev, V.G., Vastel, C., Vujčić, V., Wakelam, V., Walton, N.A., Zeippen, C., Zwölf, C.M.: A Decade with VAMDC: Results and Ambitions. Atoms **8**(4), 76 (2020). https://doi.org/10.3390/atoms8040076

[4] Moreau, N., Zwolf, C.-M., Ba, Y.-A., Richard, C., Boudon, V., Dubernet, M.-L.: The VAMDC portal as a major enabler of atomic and molecular data citation. Galaxies **6**(4), 105 (2018). https://doi.org/10.3390/galaxies6040105

[5] Regandell, S., Marquart, T., Piskunov, N.: Inside a VAMDC data node—putting standards into practical software. Physica Scripta **93**(3), 035001 (2018). https://doi.org/10.1088/1402-4896/aaa268

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

[7] Walton, N.: Meeting the user science challenge for a virtual universe. In: Toward an International Virtual Observatory, pp. 187–192. Springer. https://doi.org/10.1007/10857598_29. https://doi.org/10.1007/10857598_29

[8] Heller, S.R., McNaught, A., Pletnev, I., Stein, S., Tchekhovskoi, D.: InChI, the IUPAC international chemical identifier. Journal of Cheminformatics **7**(1) (2015). https://doi.org/10.1186/s13321-015-0068-4

[9] 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 Science Journal **18** (2019). https://doi.org/10.5334/dsj-2019-004

[10] 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., Weigel, T.: Persistent identifiers: Consolidated assertion. Research Data Alliance (2017). https://doi.org/10.15497/RDA00027. https://rd-alliance.org/system/files/PID-report_v6.1_2017-12-13_final.pdf

[11] Rauber, A., Gößwein, B., Zwölf, C.M., Schubert, C., Wörister, F., Duncan, J., Flicker, K., Zettsu, K., Meixner, K., McIntosh, L.D., Pröll, S., Miksa, T., Parsons, M.A.: Precisely and Persistently Identifying and Citing Arbitrary Subsets of Dynamic Data. Harvard Data Science Review **3**(4) (2021). https://doi.org/10.1162/99608f92.be565013. https://hdsr.mitpress.mit.edu/pub/si7wzxxa

[12] Burton, A., Aryani, A., Koers, H., Manghi, P., Bruzzo, S.L., Stocker, M., Diepenbroek, M., Schindler, U., Fenner, M.: The scholix framework for interoperability in data-literature information exchange. D-Lib Magazine **23**(1/2) (2017). https://doi.org/10.1045/january2017-burton

[13] Dubernet, M., Nenadovic, L., Doronin, N.: Spectcol: A new software to combine spectroscopic data and collisional data within vamdc. Astronomical Data Analysis Software and Systems XXI **461**, 335 (2012)

[14] Möller, T., Endres, C., Schilke, P.: eXtended CASA line analysis software suite (XCLASS). Astronomy & Astrophysics **598**, 7 (2017). https://doi.org/10.1051/0004-6361/201527203

[15] Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., Mons, B.: The FAIR guiding principles for scientific data management and stewardship. Scientific Data **3**(1) (2016). https://doi.org/10.1038/sdata.2016.18

[16] Bahim, C., Casorrán-Amilburu, C., Dekkers, M., Herczog, E., Loozen, N., Repanas, K., Russell, K., Stall, S.: The FAIR data maturity model: An approach to harmonise FAIR assessments. Data Science Journal **19** (2020). https://doi.org/10.5334/dsj-2020-041

[17] Devaraju, A., Huber, R., Mokrane, M., Herterich, P., Cepinskas, L., de Vries, J., L'Hours, H., Davidson, J., White, A.: FAIRsFAIR Data Object Assessment Metrics. Zenodo (2022). https://doi.org/10.5281/zenodo.6461229. https://zenodo.org/record/6461229#.Y85Oky_pMeY

[18] Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B., Wilkinson, M.D.: Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the european open science cloud. Information Services & Use **37**(1), 49–56 (2017). https://doi.org/10.3233/isu-170824

[19] Steinhoff, W., Tykhonov, V., Aguilar, F.: FAIR Metrics SKOS Mapping. https://doi.org/10.5281/zenodo.7113227. EOSC-synergy receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 857647. https://zenodo.org/record/7113227#.Y85KOS_pMeY

[20] Zwölf, C.M., Rixon, G.: Authentication, Authorisation and Accounting strategy. Zenodo (2015). https://doi.org/10.5281/zenodo.3936606. https://zenodo.org/record/3936606

[21] Servillat, M., Bonnarel, F., Boisson, C., Louys, M., Ruiz, J.E., Sanguillon, M.: Towards a Provenance Management System for Astronomical Observatories. Springer (2021). https://doi.org/10.1007/978-3-030-80960-7_20. https://link.springer.com/chapter/10.1007/978-3-030-80960-7_20

[22] Goble, C., Cohen-Boulakia, S., Soiland-Reyes, S., Garijo, D., Gil, Y., Crusoe, M.R., Peters, K., Schober, D.: Fair computational workflows. Data Intelligence **2**(1-2), 108–121 (2020). https://doi.org/10.1162/dint_a_00033

[23] Wilkinson, M.D., Dumontier, M., Sansone, S.-A., da Silva Santos, L.O.B., Prieto, M., Batista, D., McQuilton, P., Kuhn, T., Rocca-Serra, P., Crosas, M., Schultes, E.: Evaluating FAIR maturity through a scalable, automated, community-governed framework. Scientific Data **6**(1) (2019). https://doi.org/10.1038/s41597-019-0184-5

[24] Devaraju, A., Mokrane, M., Cepinskas, L., Huber, R., Herterich, P., de Vries, J., Akerman, V., L'Hours, H., Davidson, J., Diepenbroek, M.: From conceptualization to implementation: FAIR assessment of research data objects. Data Science Journal **20** (2021). https://doi.org/10.5334/dsj-2021-004

[25] Sun, C., Emonet, V., Dumontier, M.: A comprehensive comparison of automated FAIRness evaluation tools, pp. 44–53. http://ceur-ws.org/Vol-3127/#paper-6

[26] Masson, A., Marchi, G.D., Merin, B., Sarmiento, M.H., Wenzel, D.L., Martinez, B.: Google dataset search and DOI for data in the ESA space science archives. Advances in Space Research **67**(8), 2504–2516 (2021). https://doi.org/10.1016/j.asr.2021.01.035

[27] Guha, R.V., Brickley, D., MacBeth, S.: Schema.org: Evolution of structured data on the web. Queue **13**(9), 10–37 (2015). https://doi.org/10.1145/2857274.2857276

[28] Sansone, S.-A., , McQuilton, P., Rocca-Serra, P., Gonzalez-Beltran, A., Izzo, M., Lister, A.L., Thurston, M.: FAIRsharing as a community

approach to standards, repositories and policies. Nature Biotechnology **37**(4), 358–367 (2019). https://doi.org/10.1038/s41587-019-0080-8

[29] Ba, Y.-A., Dubernet, M.-L., Moreau, N., Zwölf, C.M.: Basecol2020 new technical design. Atoms **8**(4) (2020). https://doi.org/10.3390/atoms8040069

[30] Lannom, L., Broeder, D., Manepalli, G.: RDA Data Type Registries Working Group Output. Zenodo (2015). https://doi.org/10.15497/A5BCD108-ECC4-41BE-91A7-20112FF77458. https://zenodo.org/record/1406127/