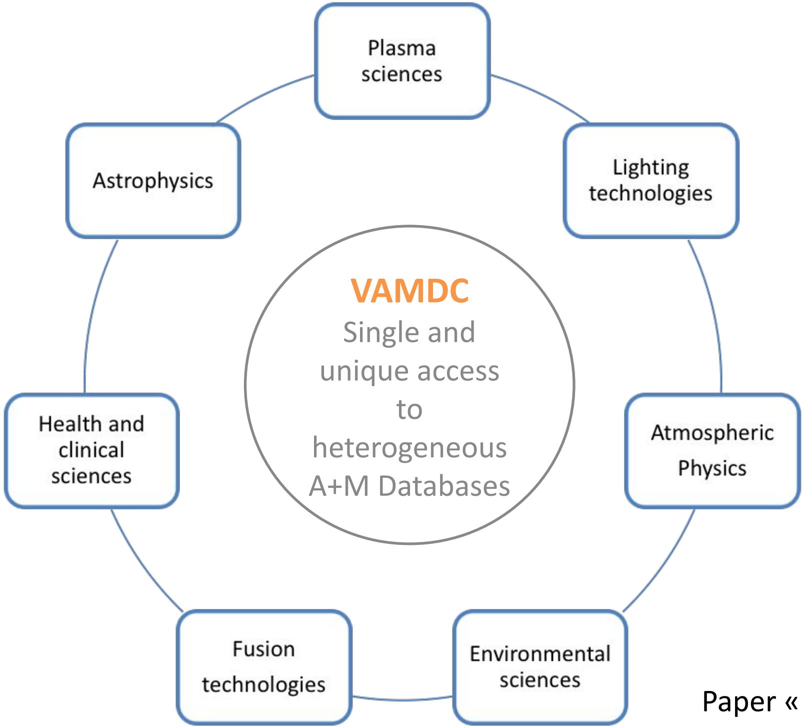
FAIR Assessment of the VAMDC infrastructure

ASOS 2023 - Paris 2023/07/13

https://tinyurl.com/asos2023

Carlo Maria Zwölf, Nicolas Moreau, Marie-Lise Dubernet

The Virtual Atomic and Molecular Data Centre in a nutshell



- E-infrastructure connecting about 40 heterogeneous databases that can be accessed from http://portal.vamdc.org/ or any VAMDC compatible tools
- Consortium of 25 partners
- High quality scientific data come from different Physical/Chemical Communities
- Provides a large dissemination platform to data producers

Paper « A decade with VAMDC : results and ambition, Atoms, 2020 »

http://dx.doi.org/10.3390/atoms8040076

List of interconnected databases

Databases	Type of A&M Data	Partners	Application's Fields
NIFS AMDIS IONIZATION	Electron-impact ionization cross-sections and rate coefficients (atoms & atomic ions)	National Institute for Fusion Science, Toki, Japan, I. Murakami	Stellar, Solar, plasma, fusion
VALD	Atomic Linelists	Uppsala, Vienna, Moscow – N. Piskunov	Stellar –Solar
NIST Atomic Spectra	Spectroscopy of Atoms –	NIST – Yuri Ralchenko	Stellar – ISM -
CHIANTI	Atomic Linelists and collisions	Cambridge (UK)+MSSL/UCL – H. Mason, G. Rixon	Solar Physics
Spectr-W3	Atomic Linelists and Collisions	Russia (RFNC VNIITF) – P. Loboda	Solar/Stellar Physics + Fusion, plasma
Stark-B	Atomic LineShifts/Broadening with charged perturbers	Observatory of Belgrade (Serbia) + Observatory of Paris (LERMA) – M. Dimitrijevic/S. Sahal-Bréchot	Stellar Physics + Plasmas
TipBase, TopBase	Atomic Linelists and Collisions from Opacity Project and IRON Project	Observatory of Paris (LERMA) + CDS (Strasbourg, Fce) – F. Delahaye/C. Zeippen/C. Mendoza	Stellar, Solar Physics,
SESAM	Electronic Spectra of atoms and molecules	Paris Obs. – E. Roueff	ISM - Stellar

List of interconnected databases

Databases	Type of A&M Data	Partners	Application's Fields
MOLD	Photo-Dissociation Cross-sections	Institute of Physics, Astronomical Obs, Belgrade, Serbia- Vladimir Sreckovic, V. Vujcic, D. Jevremovic	Stellar
BEAM-DB	Molecular/atom—electron collisions	Institute of Physics, Belgrade, Serbia Bratislav Marinkovi\'c	plasma, radiation damage
IDEABD	Dissociative electron attachment upon interaction of low energy electrons with molecules.	Innsbrück F. Duensing	Planets, ExoPlanets, ISM, Radiation Damage
AMBDAS	Collisions in plasmas (bibliographic) – searchable via processes ans species	IAEA, Vienna, Austria – C. Hill	Nuclear Fusion

List of interconnected databases

Databases	Type of A&M Data	Partners	Application's Fields
CDMS	Molecular Linelists (mm, Sub-mm)	Cologne (Germany) – S. Schlemmer	ISM + Earth+ CO
JPL	Molecular Linelists (mm, Sub-mm)	Pasadena (USA) + Cologne (Germany) – B. Drouin	ISM + Earth+CO
HITRAN	Molecular Linelists and Broadening Coefficients	Harvard (USA) + UCL – I. Gordon + L. Rothman	Earth, Planets, Exo-Planets
S&MPO	O_3 linelists	Reims (France)+ Tomsk (Russia) – V. Tyuterev	Earth – Exo-Planets
MeCaSDa	Linelists CH_4	Dijon (France) – V. Boudon	Earth, Planets, Exo-Planets, Brown dwarfs
SHeCaSDa	Sulfur Hexafluoride Calculated LInelists	Dijon – V. Boudon	Earth
TFMeCaSDa	Tetrafluoro-Methane calculated linelists	Dijon – V. Boudon	Earth
ECaSDa	Ethene Calculated LInelists	Reims – L. Daumont	Earth and Planets
GeCaSDa	GeH_4 Linelists	Dijon – V. Boudon	Planets

List of interconnected databases

Databases	Type of A&M Data	Partners	Application's Fields
RuCaSDa	RuO_4 Linelists	Dijon – V. Boudon	Nuclear Industry
TFSiCaSDa	SiF_4 Linelists	Dijon – V. Boudon	Earth
UHeCaSDa	UF_6 Linelists	Dijon – V. Boudon	Nuclear Industry
CDSD-296	CO_2 Linelists (intensity cut-off)	IAO, Tomsk – V. Perevalov	Earth, Planets, Brown Dwarfs
CDSD-1000	CO_2 Linelists (intensity cut-off)	IAO, Tomsk – V. Perevalov	Earth, Planets, Brown Dwarfs
CDSD-4000	CO_2 Linelists (intensity cut-off)	IAO, Tomsk – V. Perevalov	Earth, Planets, Brown Dwarfs
NOSD-1000	N_2O Linelists (intensity cut-off)	IAO, Tomsk – V. Perevalov	Earth, Planets
NDSD-1000	NO_2 Linelists (intensity cut-off)	IAO, Tomsk – V. Perevalov	Earth, Planets
ASD-1000	C_2H_2 Linelists (intensity cut-off)	IAO, Tomsk – V. Perevalov	Earth, Planets

List of interconnected databases

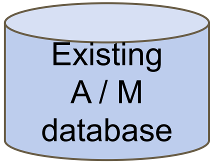
VAMDC consortium

Databases	Type of A&M Data	Partners	Application's Fields
PAH	PAH Theoretical Data and soon experimental Data	Observatory of Cagliari (Italy) – IRAP (Toulouse, France) – G. Mulas+C. Joblin	ISM, Planets, Earth
KIDA	Kinetic Data	Bordeaux (France) – P. Gratier & V. Wakelam	ISM - Planets
UdfA	Kinetic Data (ex-UMIST)	Belfast (UK) – T. Millar	ISM - Planets
BASECOL	Low Energy Molecular Collisions	Observatory of Paris – M.L. Dubernet	ISM - CO
LASP	Solid Spectroscopy Data	Obs. of Catania – G. Leto	Planets, ISM
GhoSST	Solid Spectroscopy Data	Grenoble (France) – B. Schmitt	Planets, ISM
W@DIS	Water Information System	IAO, Tomsk – A. Fazliev	Earth and Planets

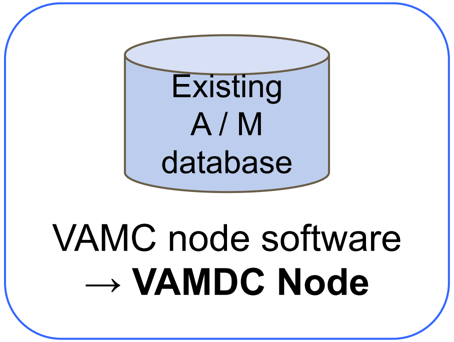
To be connected to VAMDC infrastructure

Databases	Type of A&M Data	Partners	Application's Fields
ExoMolOP	Molecular Opacities	University College London, UK – J. Tennyson	Exo, Brown Dwarf, Earth, Stellar
SSHADE	Solid Spectroscopy Data - Interface to infrastructure	Grenoble (France) & other countries – B. Schmitt et al	Earth, Comets, Exo-Planets, ISM, Planets
IAMDB	Indian Atomic and Molecular Database (atomic collisions, A+M spectroscopy)	B. Antony- Indian Institute of Technology, Dhanbad, India E. Krishnakumar - Raman Research Institute, Bengalore, India	Astrophysics, Other
DESIRE	Spectroscopy of sixth row elements (Z=72-86)	Mons University and Liege University, Belgium – P.Quinet, P. Palmeri	Plasmas – Stellar - Solar
DREAM	Radiative data for rare earth	Mons University and Liege University, Belgium – P Quinet, P. Palmeri	Stellar-Solar-Plasmas – Lighting -
PEARL	Atomic Processes	Nuclear data Center, KAERI, Daejon, South Korea Kwon Duck-Hee	Stellar-Solar-Plasmas – Fusion
Clusters	Cluster size distributions, condensation	Innsbrück F. Duensing, P. Scheier	Planets, ExoPlanets, Solvation, Biology
Additional NIFS Databases	Atomic/Molecular processes	National Institute for Fusion Science, Toki, Japan, I. Murakami	Stellar, Solar, plasma, fusion

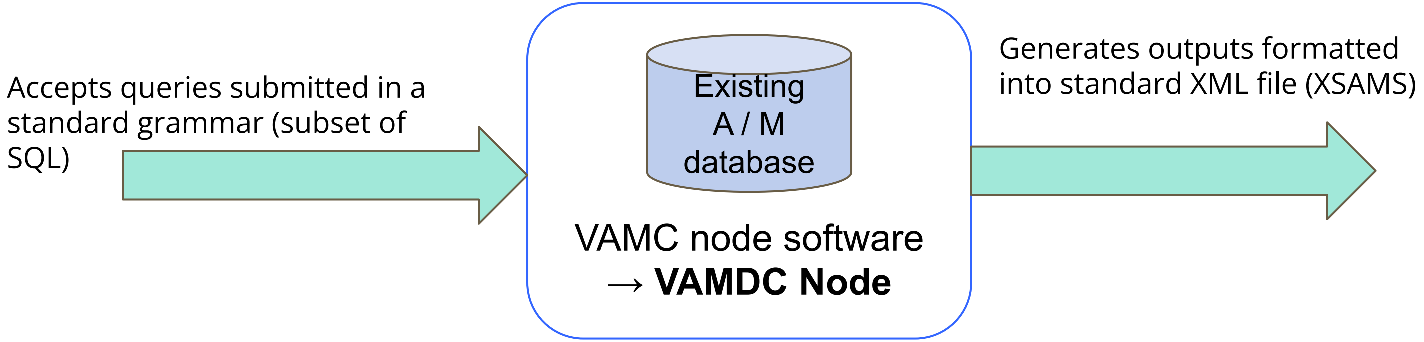
The infrastructure technical architecture



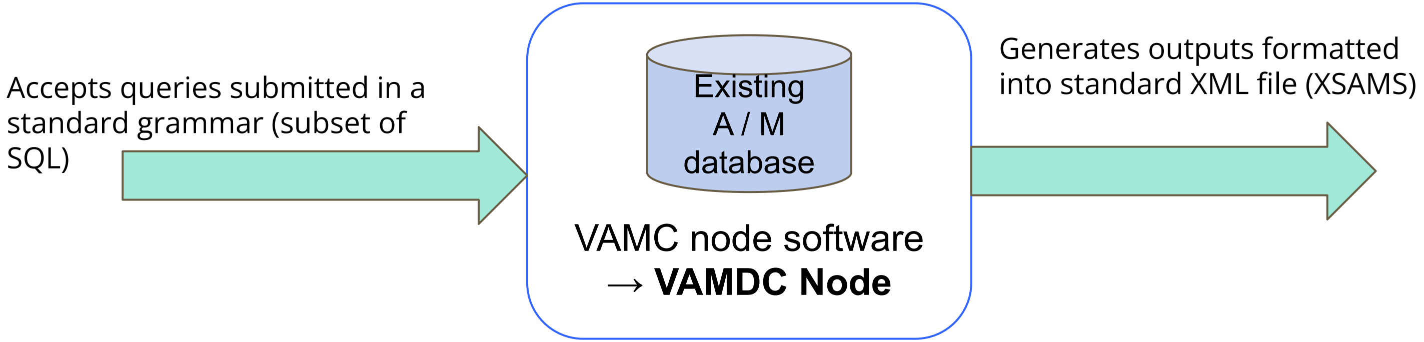
The infrastructure technical architecture



The infrastructure technical architecture

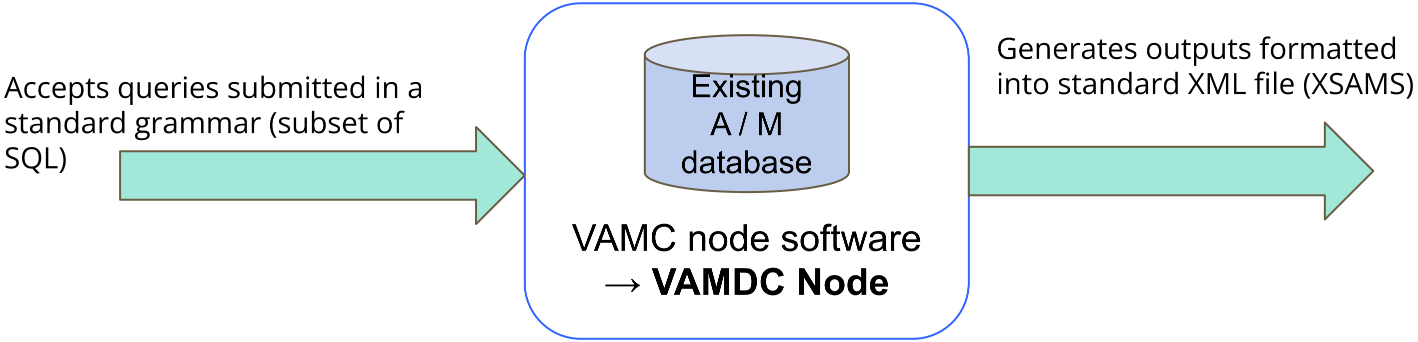


The infrastructure technical architecture

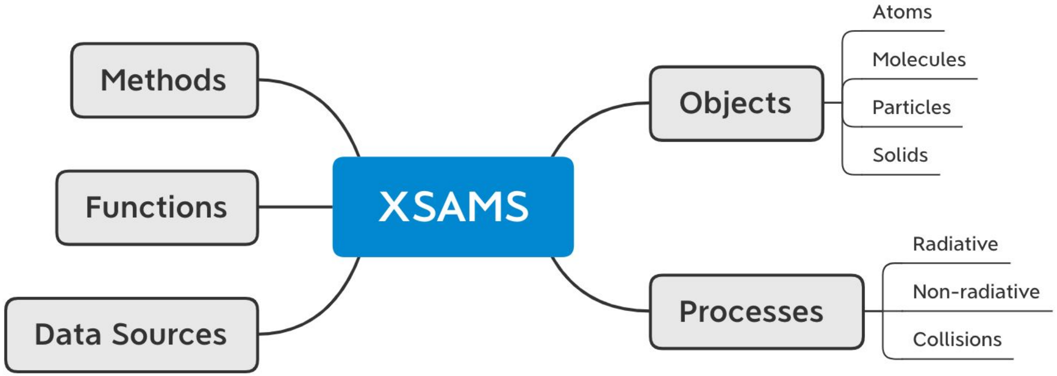


https://standards.vamdc.eu

The infrastructure technical architecture

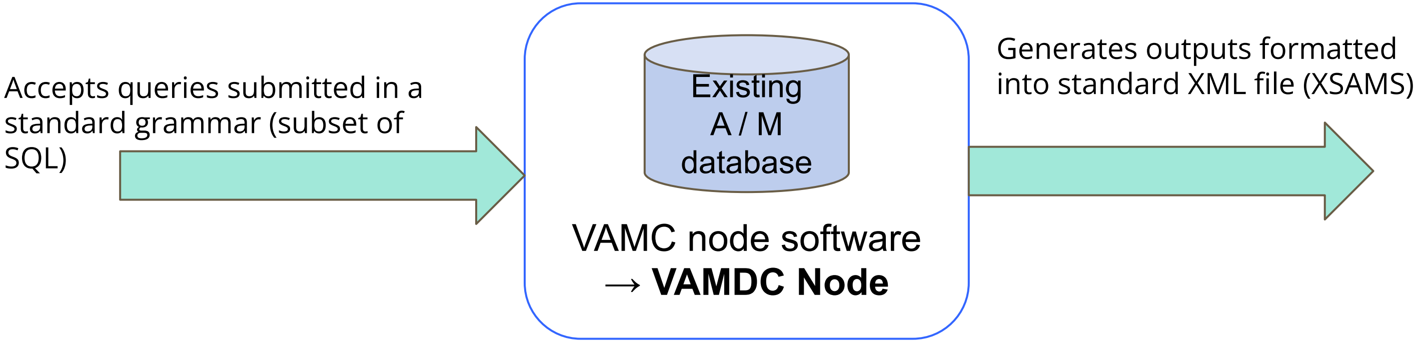


https://standards.vamdc.eu

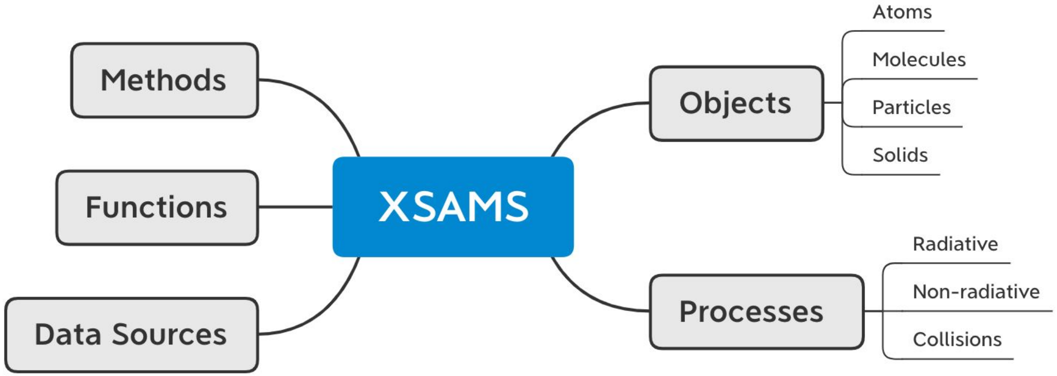


The infrastructure technical architecture

VAMDC consortium



https://standards.vamdc.eu



D.R. Schultz, ORNL; E. Roueff, ML Dubernet, N. Moreau : Observatoire Paris; S. Gagarin, P.A. Loboda, VNIITF

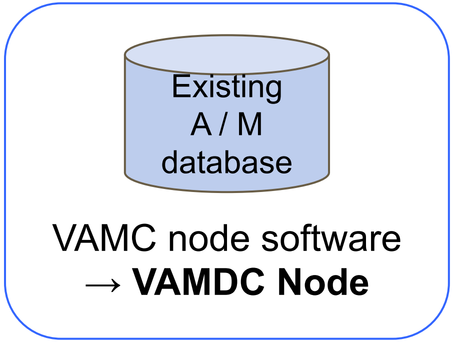
NIST National Institute of Standards and Technology U.S. Department of Commerce

Y.Ralchenko

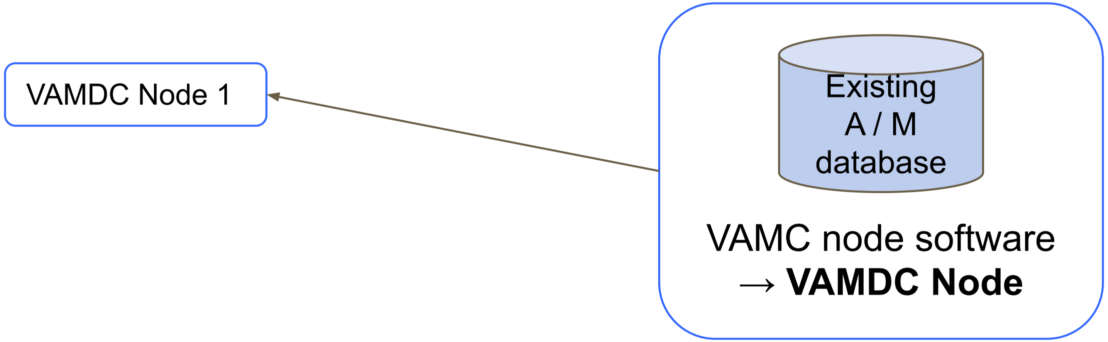
IAEA International Atomic Energy Agency

R.E.H. Clark, D. Humbert B. Braams

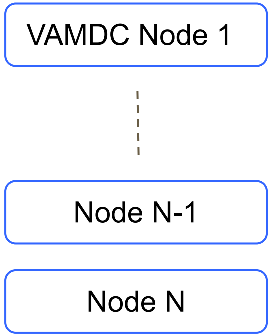
The infrastructure technical architecture



The infrastructure technical architecture



The infrastructure technical architecture



The infrastructure technical architecture

VAMDC consortium

VAMDC Node 1

Node N-1

Node N

Registries
https://registry.vamdc.org

The infrastructure technical architecture

VAMDC Node 1

Node N-1

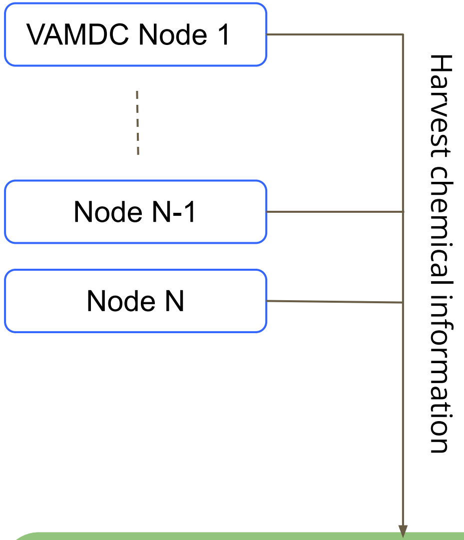
Node N

registered into

Registries
https://registry.vamdc.org

- Node unique identifier
- Scientific and tech. maintainers
- Node capabilities
- Version of implemented standard & protocols

The infrastructure technical architecture



Registries
https://registry.vamdc.org

Species Data-base
https://species.vamdc.org

For each node in VAMDC: list of species with:

- Name (in natural language)
- Stoichiometric formula
- Formula
- InChi and InChiKey
- Mass number
- Charge

The infrastructure technical architecture

VAMDC Node 1

Node N-1

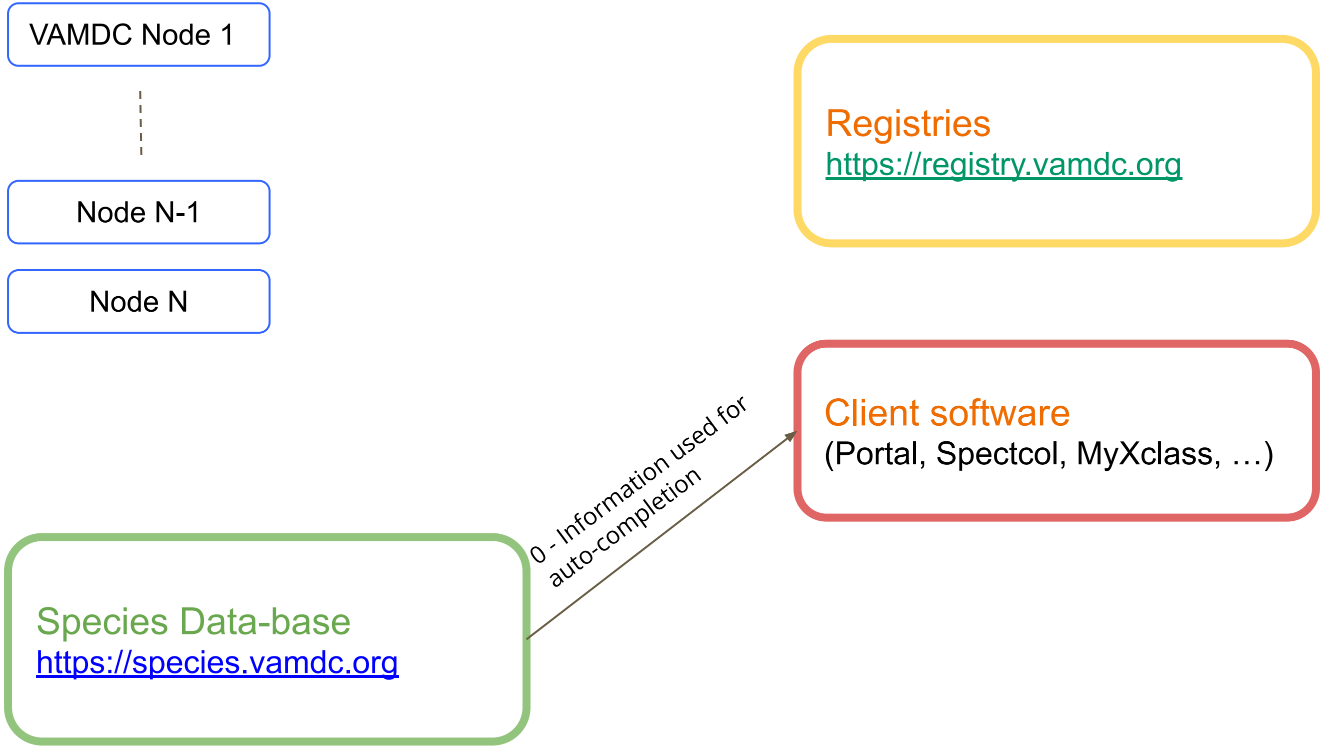
Node N

Registries
https://registry.vamdc.org

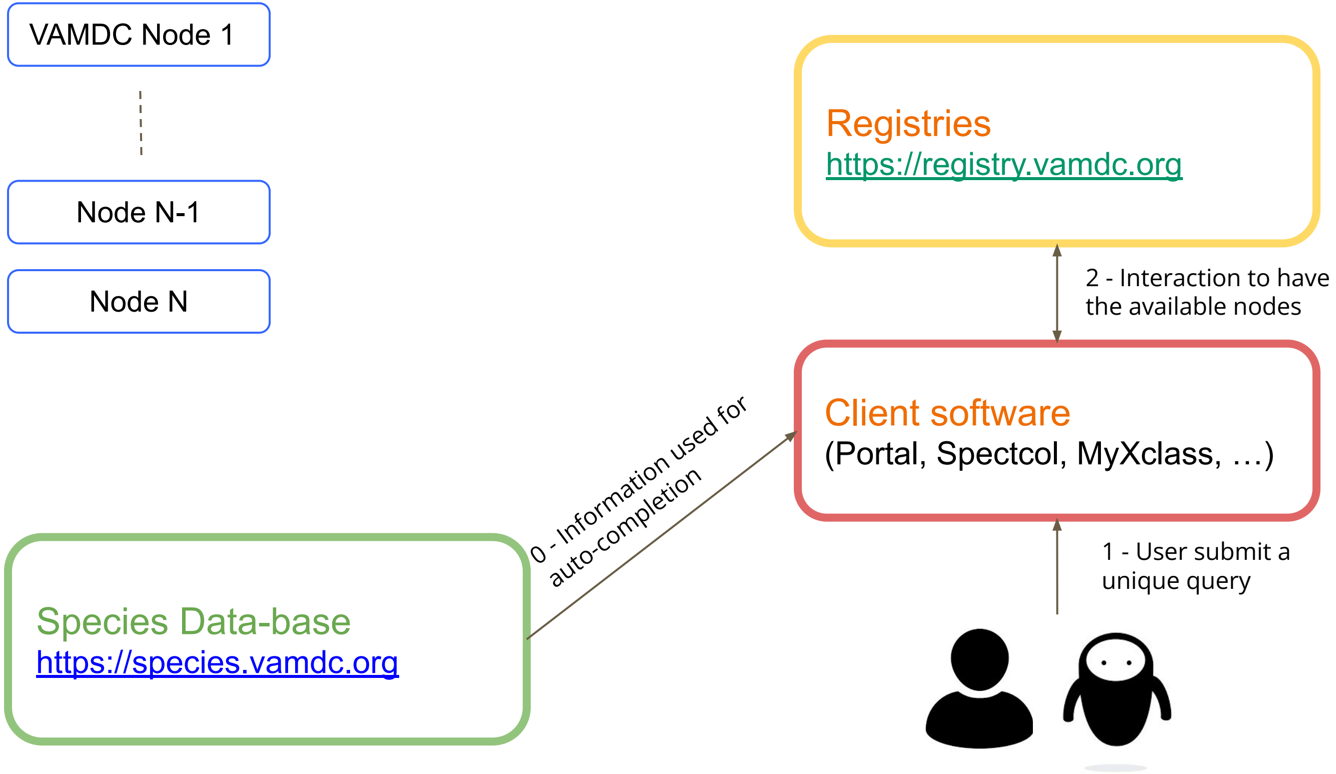
Client software
(Portal, Spectcol, MyXclass, …)

Species Data-base
https://species.vamdc.org

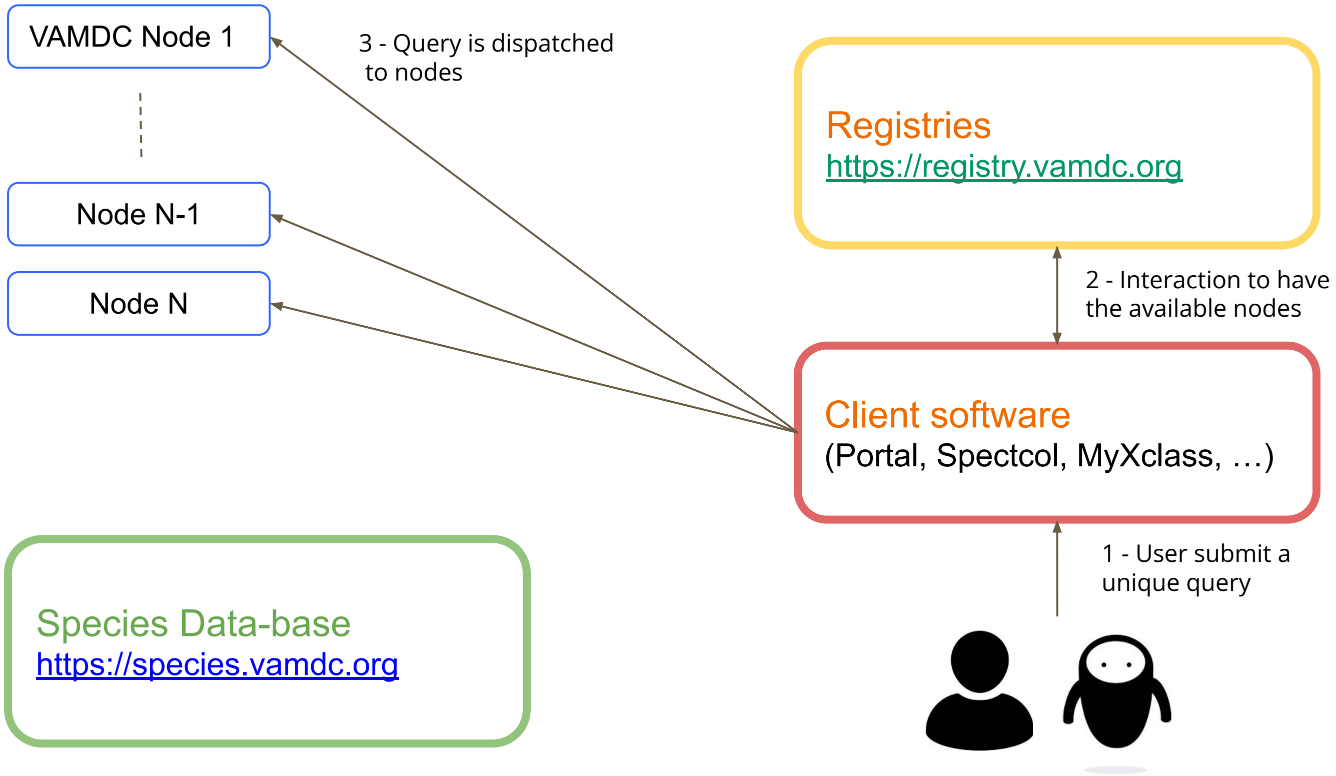
The infrastructure technical architecture



The infrastructure technical architecture

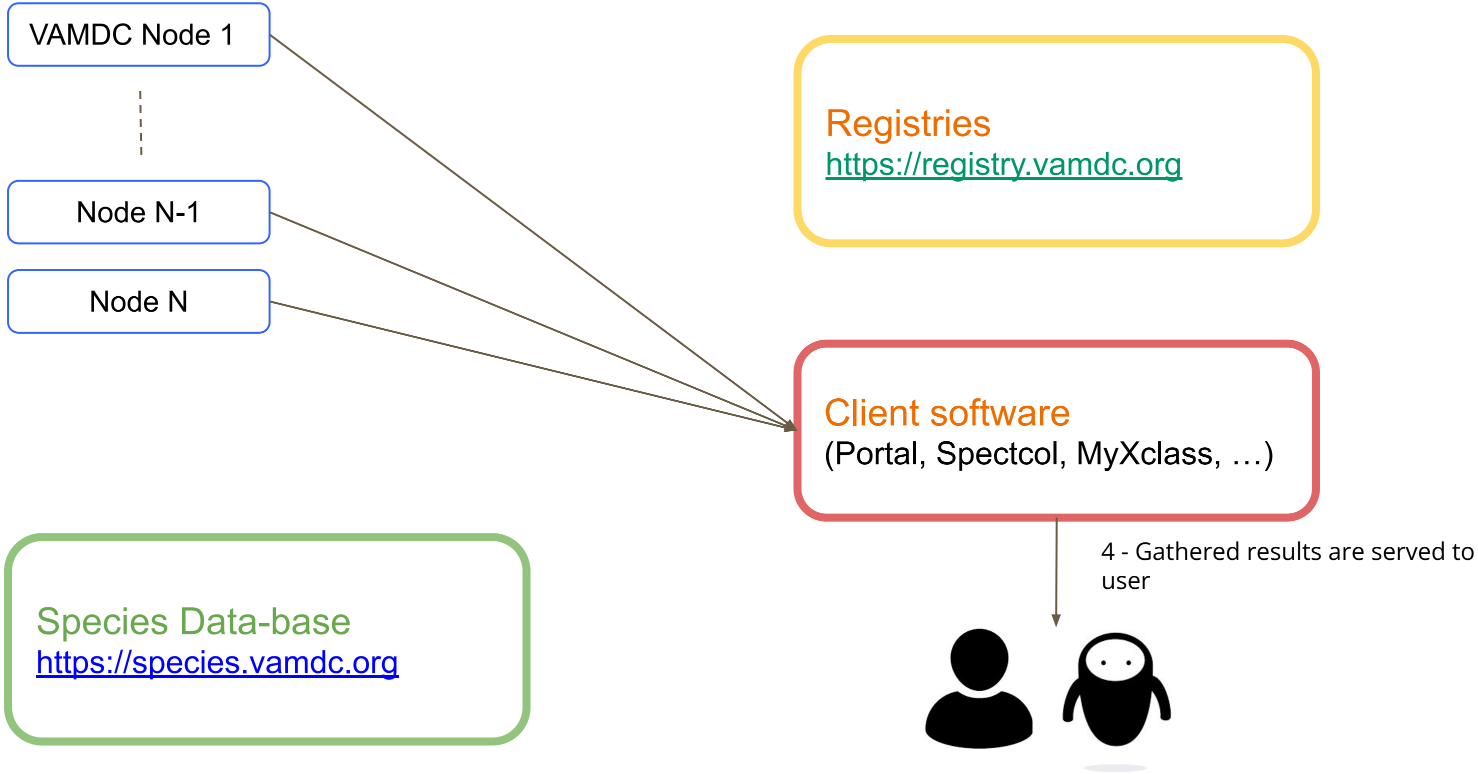


The infrastructure technical architecture



The infrastructure technical architecture

VAMDC consortium



The infrastructure technical architecture

VAMDC consortium

VAMDC Node 1

Node N-1

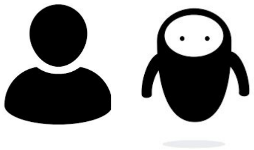
Node N

5 - details about
each served query
feed the Query Store

Registries
https://registry.vamdc.org

Client software
(Portal, Spectcol, MyXclass, …)

Species Data-base
https://species.vamdc.org



Query Store
https://cite.vamdc.org

The infrastructure technical architecture

VAMDC Node 1

Node N-1

Node N

Registries
https://registry.vamdc.org

5 - details about each served query feed the Query Store

A central repository for all the queries served by the infrastructure:

- DOI and PID for each query
- Original served query
- Version of the node
- Version of the used standards
- List of publications needed to compute the results
- Direct access to the produced output
- When the data DOI is cited all the producers receives credits

Query Store
https://cite.vamdc.org

Species Data-base
https://species.vamdc.org

The infrastructure technical architecture

VAMDC consortium

Interconnected nodes

Registries

Species data-base

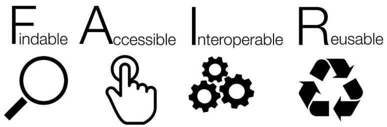
Query Store

This infrastructure has been designed since 2009 so that data are:

- Easy to discover and to find
- Immediately accessible
- Interoperable
- Easy to reuse (mechanisms for data-citation and delegation of bibliographic credits)

This is Findable Accessible Interoperable Reusable ante litteram!

The Fair Principles: definitions and impacts



First definition : Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Findable

- **F1**. (Meta)data are assigned a globally unique and persistent identifier
- **F2**. Data are described with rich metadata
- **F3**. Metadata clearly and explicitly include the identifier of the data they describe
- **F4**. (Meta)data are registered or indexed in a searchable resource

Accessible

- **A1**. (Meta)data are retrievable by their identifier using a standardised communications protocol
  - **A1.1** The protocol is open, free, and universally implementable
  - **A1.2** The protocol allows for an authentication and authorisation procedure, where necessary
- **A2**. Metadata are accessible, even when the data are no longer available

Interoperable

- **I1**. (Meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- **I2**. (Meta)data use vocabularies that follow FAIR principles
- **I3**. (Meta)data include qualified references to other (meta)data

Re-usable

- **R1**. (Meta)data are richly described with a plurality of accurate and relevant attributes
- **R1.1**. (Meta)data are released with a clear and accessible data usage license
- **R1.2**. (Meta)data are associated with detailed provenance
- **R1.3**. (Meta)data meet domain-relevant community standards

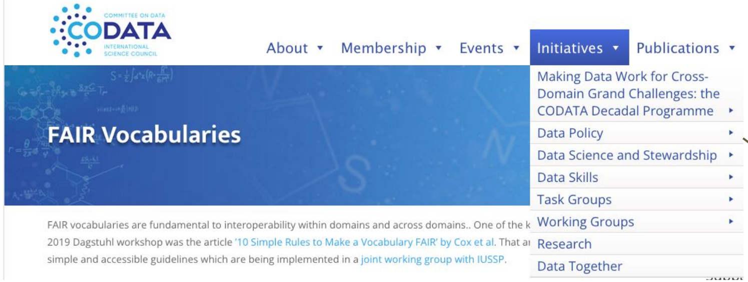
The Fair Principles: definitions and impacts

VAMDC consortium

Findable Accessible Interoperable Reusable

First definition : Wilkinson et al, The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

In few years FAIR became a MUST for data oriented initiatives and communities



IUPAC INTERNATIONAL UNION OF PURE AND APPLIED CHEMISTRY

WP : Chemistry
https://iupac.org/worldfair-global-cooperation-on-fair-data-policy-and-practice/

EU Framework program

https://www.rd-alliance.org
Working Groups that deliver recommendations
Interest Groups
Community of Practices

GO FAIR
FAIR Principles Implementation Networks
https://www.go-fair.org

The Fair Principles: definitions and impacts

VAMDC consortium

FAIR Data Maturity Model WG

WG

Taxonomy:

Posts | Wiki | Events | Repository | Outputs | Case Statements | Plenaries | Members

Group Status: WGs Maintaining deliverables (maintenance group)

Status: Recognised & Endorsed
Chair (s): Edit Herczog, Keith Russell, Shelley Stall
Secretariat Liaison: Stefanie Kethers
TAB Liaison: Karin Breitman

DOI: 10.15497/rda00050

FAIR Data Maturity Model: core criteria to assess the implementation level of the FAIR data principles

Webinar - The RDA FAIR Data Maturity Model WG: Aligning International Initiatives for Promoting and Assessing FAIR Data - 19/11/2020

Webinar - Connecting Data, Institutions and People: FAIR Digital Objects, RDA outputs and the design of the DiSSCo Research Infrastructure - 18 March 2021

The RDA FAIR Data Maturity Model Working Group develops as an RDA Recommendation a common set of core assessment criteria for FAIRness and a generic and expandable self-assessment model for measuring the maturity level of a dataset. The aim is not to develop yet another FAIR assessment approach but to build on existing initiatives, looking at common elements and allowing the group to identify core elements for the evaluation of FAIRness. That will increase the coherence and interoperability of existing or emerging FAIR assessment frameworks and it will ensure the combination and compatibility of their results in a meaningful way.

The WG brings together stakeholders from different scientific and research disciplines, the industry and public sector, who are active and/or interested in the FAIR data principles and in particular in assessment criteria and methodologies for evaluating their real-life uptake and implementation level.

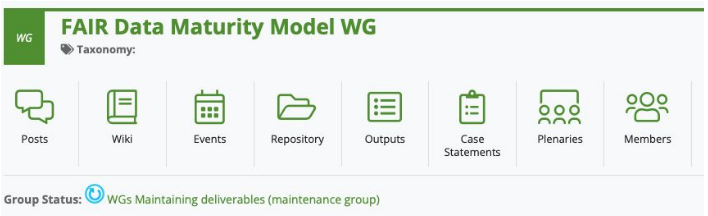
Building the social and technical bridges to enable open sharing and re-use of data

RDA RESEARCH DATA ALLIANCE

O&A Members 71

Active Organisational & Affiliate members

The RDA evaluation framework



DOI: 10.15497/rda00050

The recommendation consists of ~40 indicators

- 7 for **F**indable principle,
- 12 for **A**ccessible
- 12 for **I**nteroperable
- 10 for **R**eusable

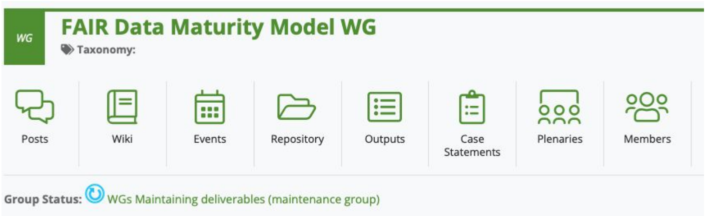
Each indicator has a priority which determines its importance

- **Essential**
- **Important**
- **Useful**

The level of maturity for the indicator is an integer taking the following values

- 0 - non applicable
- 1 - not being considered yet
- 2 - under consideration or in planning phase
- 3 - in implementation phase
- 4 - fully implemented

The RDA evaluation framework



DOI: 10.15497/rda00050

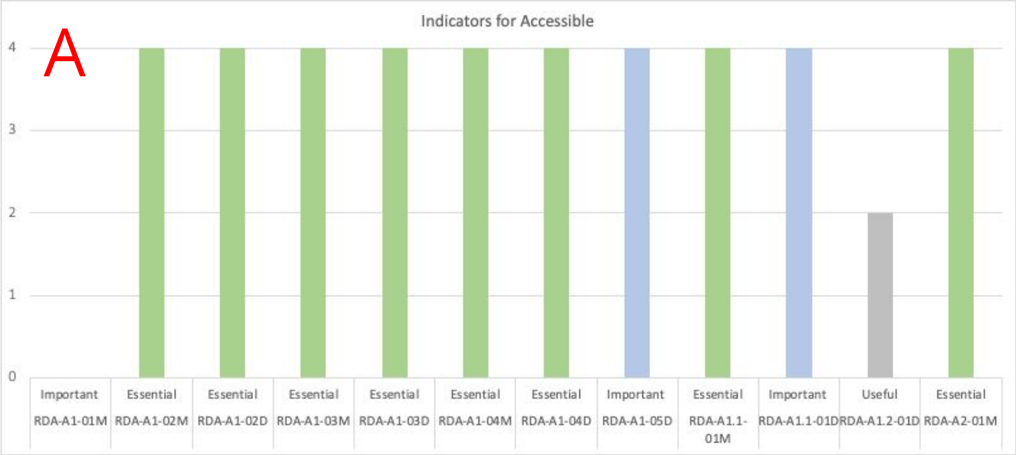
We went through all criteria to assess the FAIRness of VAMDC

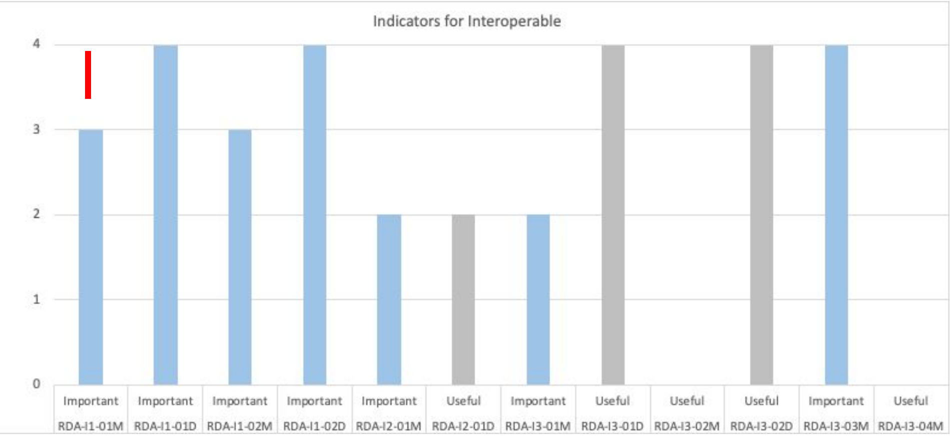
A priori thoughts:

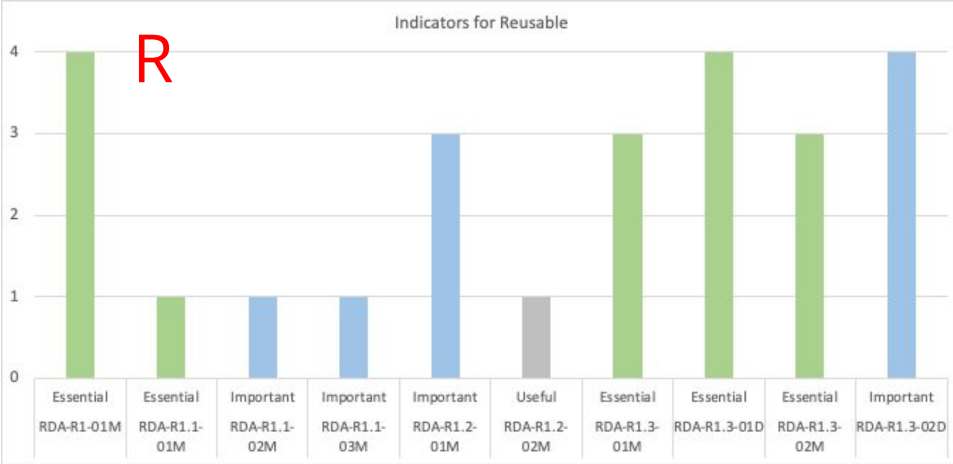
- Maybe a quite technocratic generic document. Is it relevant for our science platform?
- We are already FAIR. Do we need this kind of evaluation?

Actually we learnt a lot!

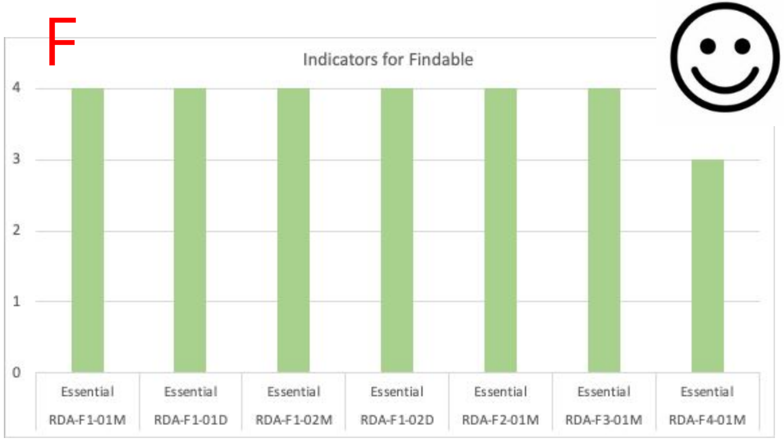
Overview of the evaluation results

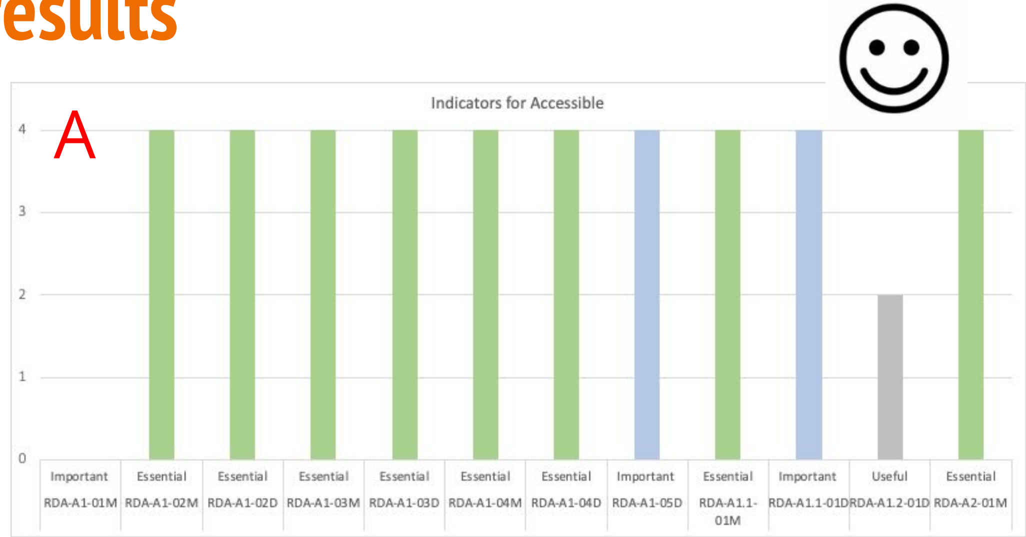


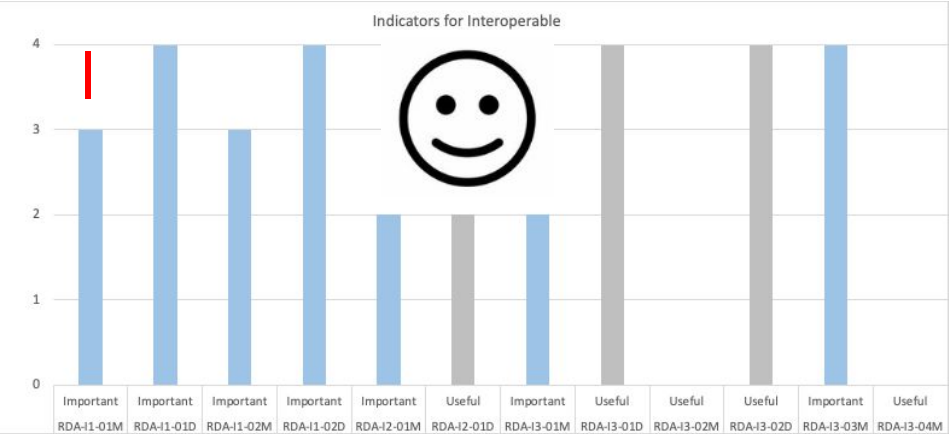


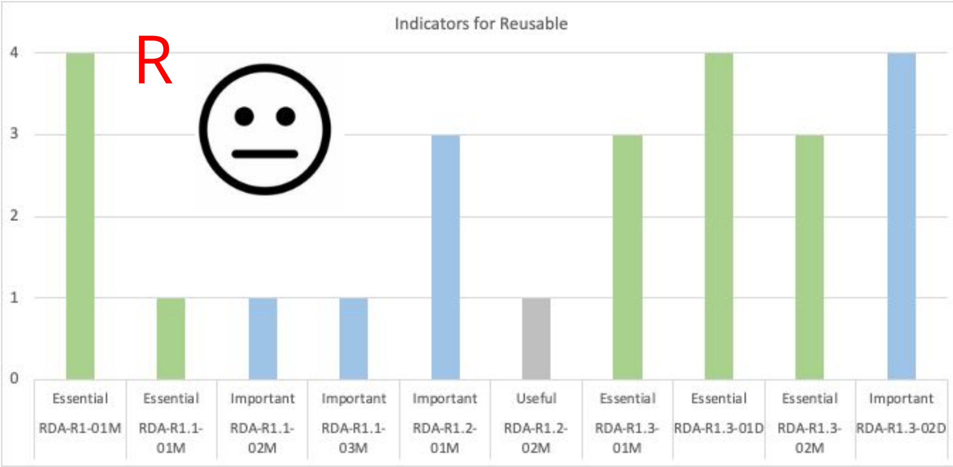


Overview of the evaluation results









What we learnt: what we may improve

Interconnected nodes

Registries

Species data-base

Query Store

What we learnt: what we may improve

Interconnected nodes

Registries

A better integration between information from registries and Species DB is encouraged. A single entry point for all the information about a given node

Species data-base

Query Store

What we learnt: what we may improve

Interconnected nodes

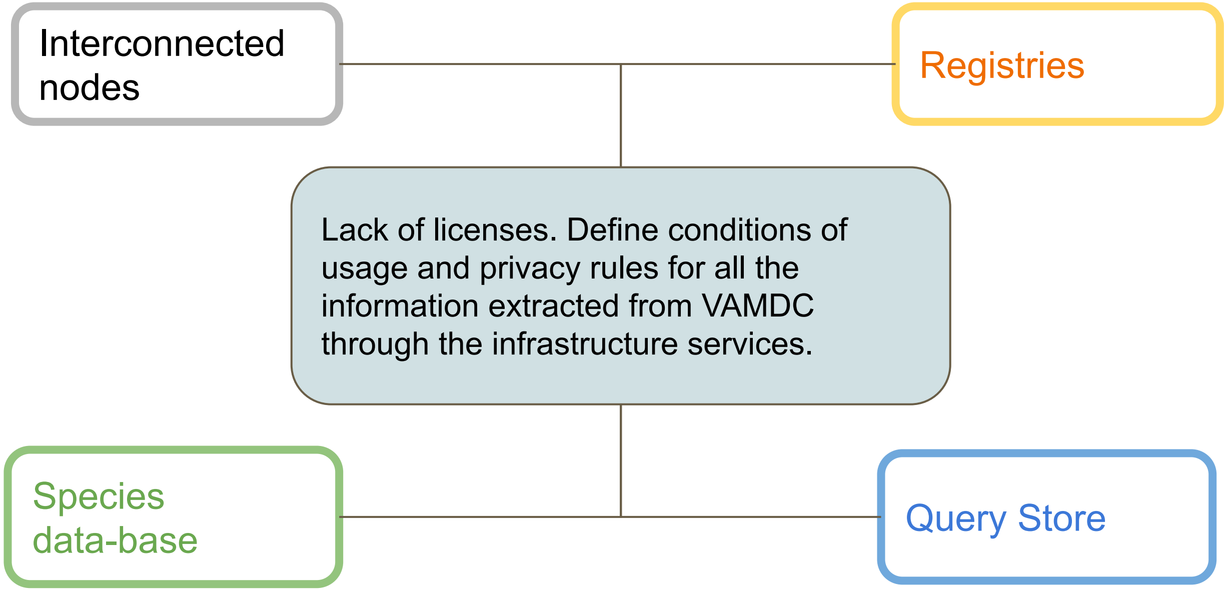
Registries

When Specied DB refers to a node element, this reference should contain a machine resolvable pointer to the node entry into the registries

Species data-base

Query Store

What we learnt: what we may improve



What we learnt: what we may improve

Interconnected nodes

Registries

Three aspects may be improved in XSAMS standard

- **Register this format as a standard into FAIRSharing “registry of types”**
- Several enumerations in XSAMS (*CategoryType* in *Methods*, *dataDescription* Values in *DataSet*) are not defined in the schema nor in external dictionaries → **Define proper machine actionable FAIR dictionaries for these terms**
- For codes of Processes **there should be a PID resolving to the definition of these codes**

Species data-base

Query Store

PID is a Persistent Resolvable unique identifier

What we learnt: what we may improve

Interconnected nodes

Registries

- The speciesDB/QS use no dictionary to express knowledge → **Define such dictionaries**
- The speciesDB/QS contain references to the nodes but not to the registry information → **have a link to the related node-entry into registries.**
- The speciesDB/QS contain references to the standards, but not in machine actionable way → **point to standards in machine actionable way**
- Information from SpeciesDB/QS do not follow particular standard/format
  - → **Introduce standard for information from SpeciesDB/QS**
  - → **Describe the interfaces of SpeciesDB/QS in a standard way (both human and machine oriented ones)**

Species data-base

Query Store

What we learnt: what we may improve

Interconnected nodes

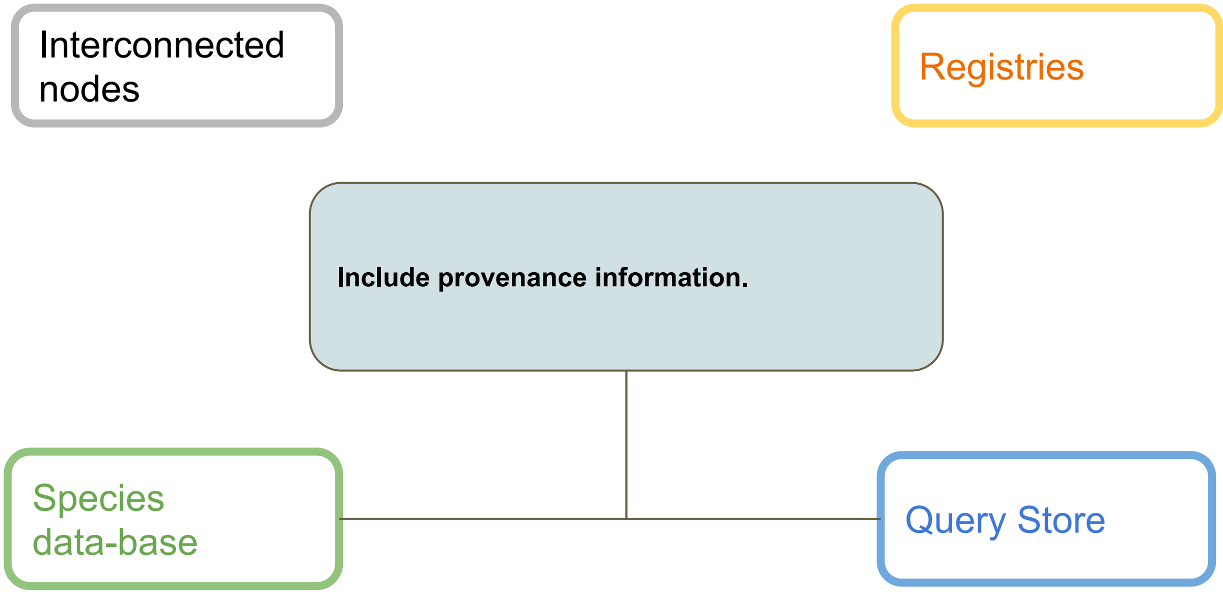
Registries

- Registries uses terms from controlled vocabularies. Definition of these terms is not accessible directly → **Use FAIR compliant vocabulary (terms is a PID to its definition).**
- The registries contain references to the standards, but not in machine actionable way → **point to standards in machine actionable way**

Species data-base

Query Store

What we learnt: what we may improve



Conclusions

The analysis was very useful !!

- 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

- The FAIRness of VAMDC
  - is satisfying for people aware of the infrastructure subtleties
  - may be improved for newcomers and/or for transdisciplinary activities

The main lines of improvement identified 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;
- The cross-references of metadata between the different pillar components 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;

Conclusions

The analysis was very useful !!

- 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

- The FAIRness of VAMDC
  - is satisfying for people aware of the infrastructure subtleties
  - may be improved for newcomers and/or for transdisciplinary activities

The main obstacles that explain the unsatisfactory evaluation for the Reusable principle are related to the absence of a license-policy and the non-adoption of a standard representation for provenance information.

Very hard to define licenses in an international (i.e. multi-juridical) framework

Standard (W3C/IVOA) provenance representations are difficult to handle for end-users. A readme file fulfils VAMDC users' needs

Conclusions

VAMDC consortium

The analysis was very useful !!

- 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

- The FAIRness of VAMDC
  - is satisfying for people aware of the infrastructure subtleties
  - may be improved for newcomers and/or for transdisciplinary activities

The main obstacles that explain the unsatisfactory evaluation for the Reusable principle are related to the absence of a license-policy and the non-adoption of a standard representation for provenance information.

Very hard to define licenses in an international (i.e. multi-juridical) framework

Standard (W3C/IVOA) provenance representations are difficult to handle for end-users. A readme file fulfils VAMDC users' needs

- To go deeper: https://doi.org/10.1140/epjd/s10053-023-00649-x - Eur. Phys. J. D77, 70 (2023)

- These slides: https://tinyurl.com/asos2023