

VAMDC Standards

Where do we want to go?

Thomas Marquart

Uppsala University, Sweden

Where are we? Standards overview

Data node API: VAMDC-TAP

- Query language: VSS2 (VAMDC SQL Subset #2)
- Dictionaries
 - Restrictables
 - Requestables
 - HTTP-Headers
- Data transfer format: **XSAMS** (XML Schema for Atoms, Molecules & Solids)
 - "Case-by-case" for molecular quantum numbers
 - InChi(Key)s
 - Units
 - (Returnables)

Processors / Consumers API

Registry

Central Logger API

+ Documentation!

Questionnaire results! #1

19 responses

- 15 data providers
- 8 users (just 2 "user only")
- Suggesting to keep the questionnaire open!

Q1: VSS2 is good enough! (17 answ.)

- Except for solids
- Implementation has flaws.

Questionnaire results! #2

Asynchronous queries for large data:

- 50% think it should be a priority.
- Political issue!

Do you know of data that do not fit into XSAMS:

- No, 78 %
- Yes:
 - Matrices
 - Molecules in solid state
 - Surface description and processes

Solids, a better S in XSAMS!

Bernard Schmitt (Grenoble) suggested a replacement for <Solids>:

"I proposed in 2012 an extension for solid molecules (for the last 'S' of XSAMS) derived from SSDM (much much simplified!) that could allow some interoperability with SSHADE (similar to VAMDC but for solids). We plan in our Europlanet program to provide an improved interface between SSHADE and VAMDC but this will be only possible if this extension is implemented."

- Sample
 - Material constituents
 - Species
 - Layers

Was not implemented then, other things had to settle first for us. Time might be ripe now.

Few nodes currently use XSAMS' <Solids>, change would impact few, but those heavily.

Would impact dictionaries, thereby VSS2, portal.

Data model structure for “solids” in XSAMS

(in violet: corresponding SSDM Keywords)

1.1 Sample Table

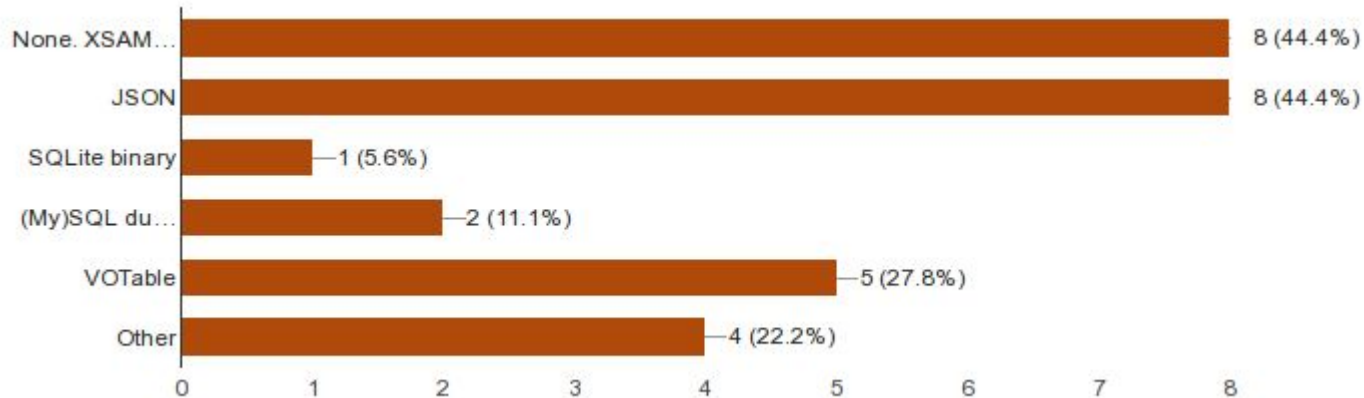
Key-word	Type	Description
SolidID <i>sample_uid</i>	string	Unique identifier code (UID) given to the sample
SolidName <i>sample_name</i>	string	Name of sample given by experimentalist
SolidRoughness <i>sample_surface_roughness_macro + _micro</i>	openum(text)	Qualitative macroscopic surface roughness of the top layer of the sample Enum: {no, polished, very low, low, medium, high, rough, unknown}
SolidThickness (+ Accuracy) <i>sample_thickness (+ _error)</i>	float [Data]	Sample thickness ➔ Calculated from the sum of the layer thicknesses ("layerThickness")
SolidThicknessUnit <i>sample_size_unit</i>	enum	Unit for the sample thickness Enum*: {nm, micron, mm, cm, m}
SolidMass <i>sample_mass</i>	float [Data]	Total mass of the sample
SolidMassUnit (g)	string	Unit of mass of the sample
SolidTemperature (+ Accuracy) <i>sample_temperature (+ _error)</i>	float	Effective temperature of the sample
SolidTemperatureUnit <i>sample_temperature_unit</i>	enum	Unit of temperature Enum*: {K, C}
SolidPressure (+ Accuracy) <i>sample_pressure (+ _error)</i>	float [Data]	Effective hydrostatic pressure of the sample
SolidPressureUnit <i>sample_pressure_unit</i>	enum	Unit of hydrostatic pressure Enum*: {Pa, hPa, mbar, bar, atm, torr}

Case-By-Case

Molecular quantum numbers.

- Suggestions as in questionnaire.
 - Agreement from several, most have no opinion or experience.
 - Comment: Not sure which "label" is being referred to with respect to C4. There is already a lpos which should be extended to allow additional, arbitrary labels.
- Further comments:
 - Nonlinear polyatomic.
 - Add a tag, which allows to label conformers.

Additional data formats



Other: CSV, HDF-5

Many of the formats are "table-based"

⇒ We need a serialization of XSAMS into limited number of tables.

⇒ Same table layout can be re-used through all formats, but one fits all nodes?

Offer via Processor services instead?

(Questionnaire: Portal & Processors)

All seems in good order and working to satisfaction!*

Few comments:

- Merging of data. (multiplexor not usable; "real" merging)
- Solids querying & data format.
- The species database (species name aliases, distinguishing molecules and isotopologues lets the portal down.

Suggested change: DataSets

New model for datasets citation and extraction reproducibility in VAMDC.

- General support, no objections.
- Make better use of DOI ?
- Suggestion: Allow technical improvements from implementation experience, only then release as standard.

Suggested change: D/T vs. H

Should Deuterium and Tritium get the letters D and T, respectively, instead of H?

- 8 votes for allowing, 5 against.
- BUT:
 - Allow in queries only, translate to H-isotope?
 - Different ElementSymbol from different nodes for same atom.
 - Nodes would have to translate the "other" version of queries.

Suggestion: To keep interoperability, either **replace** H with D/T or leave as is.

Suggested change: Relax InChi requirement

Favourably received, in general.

But:

- Require either InChi or stoichiometric formula to be present?
- How to deal with undefined species in the species database? Better to let SpeciesID of Product be "unknown" ?

Atomic States are not easily matched!

Identify the problem.

Work on a first draft for solution.

Limited resources
⇒ prioritize!

Last words

Just to repeat: to make VAMDC a reference source for A&M data we need to be able to set quality ranks, identify identical parameters and merge multiple sources according to the quality ranking.

It would be possible to implement the transformation of hyperfine splitting - calculating the set of lines and its oscillator strengths.

We need to add to the list of units "m²/sr" used for differential cross sections.

Live long and prosper!

It will be very useful to query the VAMDC databases with an id for molecule or for radiative transition in my case in order to have a link to refer.

Long live VAMDC!

I guess it is currently not unique how states (and references, ...) are restricted if not explicitly done. If a user restricts transitions via a wavelength range, will he get still all states or only those, which are necessary to describe the transitions. I would prefer to have common rules for these cases and I assume that it will improve the performance if all states are return unless there are restrictions applied for the states.

Standards release strategy for near future

Make a release (17.11 ?) this year

- Don't break anything!
- Only additions.
- No new XSAMS namespace.
- Nodes, Processors and tools do not need to be updated or re-deployed.
- No new registry.

Defer deeper changes to 2018 release.

Summary of session 1: Coming release

Issue	Timescale	Comptib	Recommendation
InChi	days	yes	Several possibilities. Decide after example from Postler.
SpeciesID Restrictable	days	yes	Add it.
Which states for transitions	-	-	Up to node to decide. No action
Deuterium & Tritium	days	no/yes	Leave as is.
Case-By-Case	days	yes	Add suggested cases.
Add process class: angular scattering	days	yes	Add it.
Unit suggestions	days	yes	Add, if new dimensionality.
Asynchronous access	weeks	yes	Allow, optional. Implement in NS
Query store	weeks	yes	Add Element for token and success
Allow https	weeks	yes?	Check standards for http-only specs

Summary of session 1: After next release

Issue	Timescale	Comptib	Recommendation
Replace Solids Schema	months	no	Iterate with B.Schmitt and Bordeaux. Implement if resources available.
DataSets	months	maybe	Implement and test, iterate once before release.
Atomic States	months	maybe	Guy gathers information, based on formalized string representation.
New DataFormats	months	(yes)	Defer, too big a task to tackle now.

InChi, possible solutions

- Make InChi non-mandatory
- Use non-xsdt validation to enforce InChi for non-products.
- Make subclass of species without inchi.
- Use "special" inchi which means "unknown".
- Leave out Species altogether, use "unknown" SpeciesID.

NodeSoftware

- Currently "rolling release" from GitHub.
- Proper release after Django 1.11 (due april, long-term support until dec 2020)

Needs resources! (currently available mini-grants?)

- VSS2-queries with prefixes poorly supported.
- Several small-to-medium sized issues.