Virtual Atomic and Molecular Data Centre project and the lessons learned

N. Piskunov

Department of Physics and Astronomy

Uppsala University

Sweden

A bit of history…

- The VAMDC project started as a general attempt to homogenize the exchange of atomic and molecular data.
- Before VAMDC (2005) we had at least 3 types of databases offering A&M data:
  1. Comma-separated format with fixed values and units as in VALD and NIST
  2. Fixed (cryptic) format as in HITRAN, CFA Kurucz etc.
  3. Binary or even more obscure formats complemented by reading routines (CHIANTI, CFA Kurucz CDs)

The VAMDC consortium

The VAMDC consortium was born out of 3 groups with initiative and experience:

- Paris with XSAMS draft (in collaboration with the NIST and the IAEA) and the concept that all A&M data should be self-documented at least when delivered to a user.
- UK/Cambridge with the VO experience and a beautiful (in its simplicity) idea to sell the VO software to the EU one more time under a different name.
- Uppsala University with the experience of VALD that was running pretty much by itself from 1997.

The VAMDC consortium

You can now try to guess which concepts ended up in VAMDC…

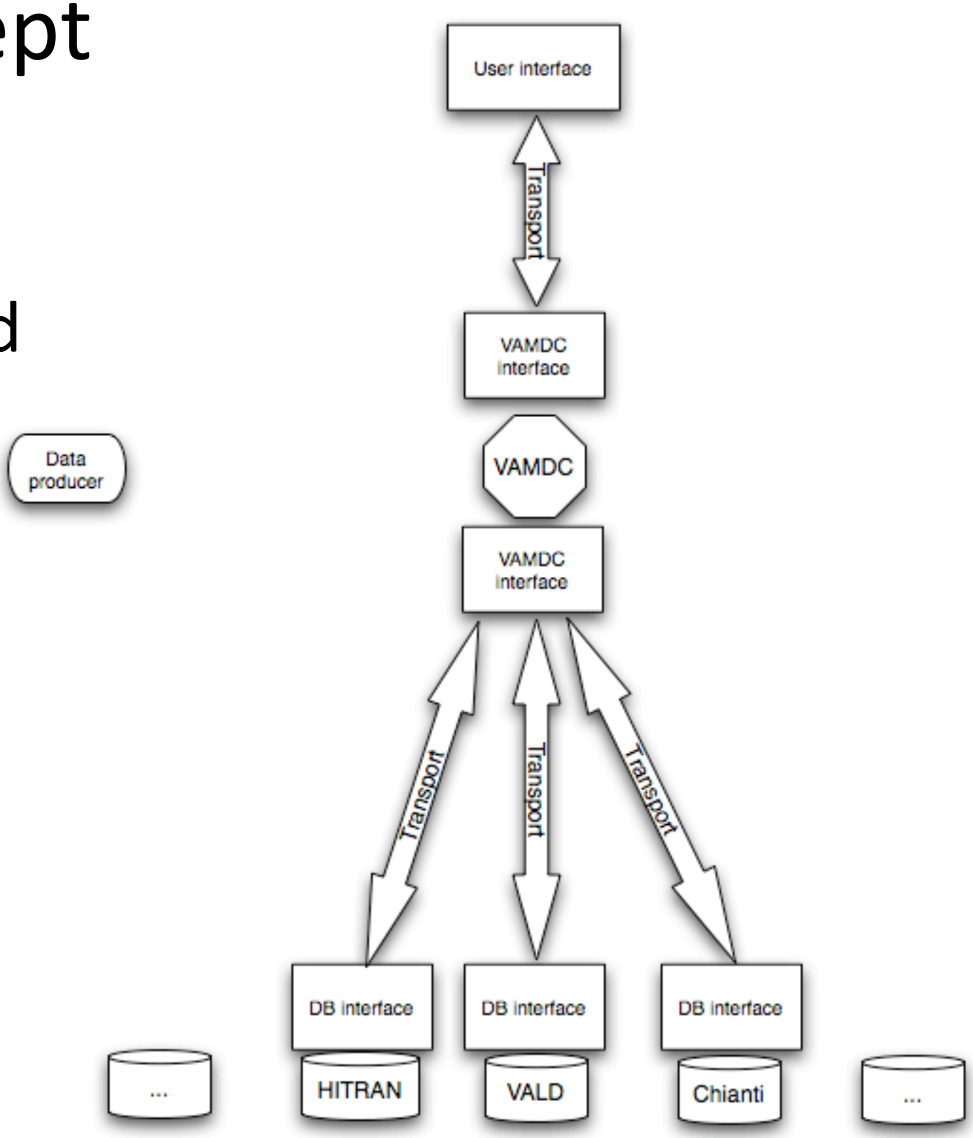
These three parties were complemented by another 13 that covered the whole spectrum from a *helping hand* to "*I'll take the money but will continue with my own business*"

In the end the VAMDC was produced by 7-8 partners!!!

EU provided large amount of money but also annoyed us with lots of paperwork.

Initial concept

- Individual nodes (databases) are accessed via a single portal.
- The data is transported in XSAMS format.
- The data from different nodes is merged into a single dataset before delivered to the user.



Working format

- Regular workshops of the Tiger Team < 10 people
- One week per workshop, changing places: Paris, Cologne, Uppsala, Cambridge, Dublin, Vienna etc.
- Each workshop dedicated to a specific task: XSAMS generators, portal, node software, registry etc.
- A workshop usually started with a brain storm about the requirements, then people worked on a various prototypes (mostly to prove their point) and finished by comparing what was done. The actual implementation was done between the workshops.

Technical aspects

- Githab was the version control system
- Java and Python were the programming languages (although others were used from time to time)
- Django framework was the base for the node software (Python version)
- Most qualified people were given the opportunity to work and not to be distracted by reports, managements etc.

VAMDC components

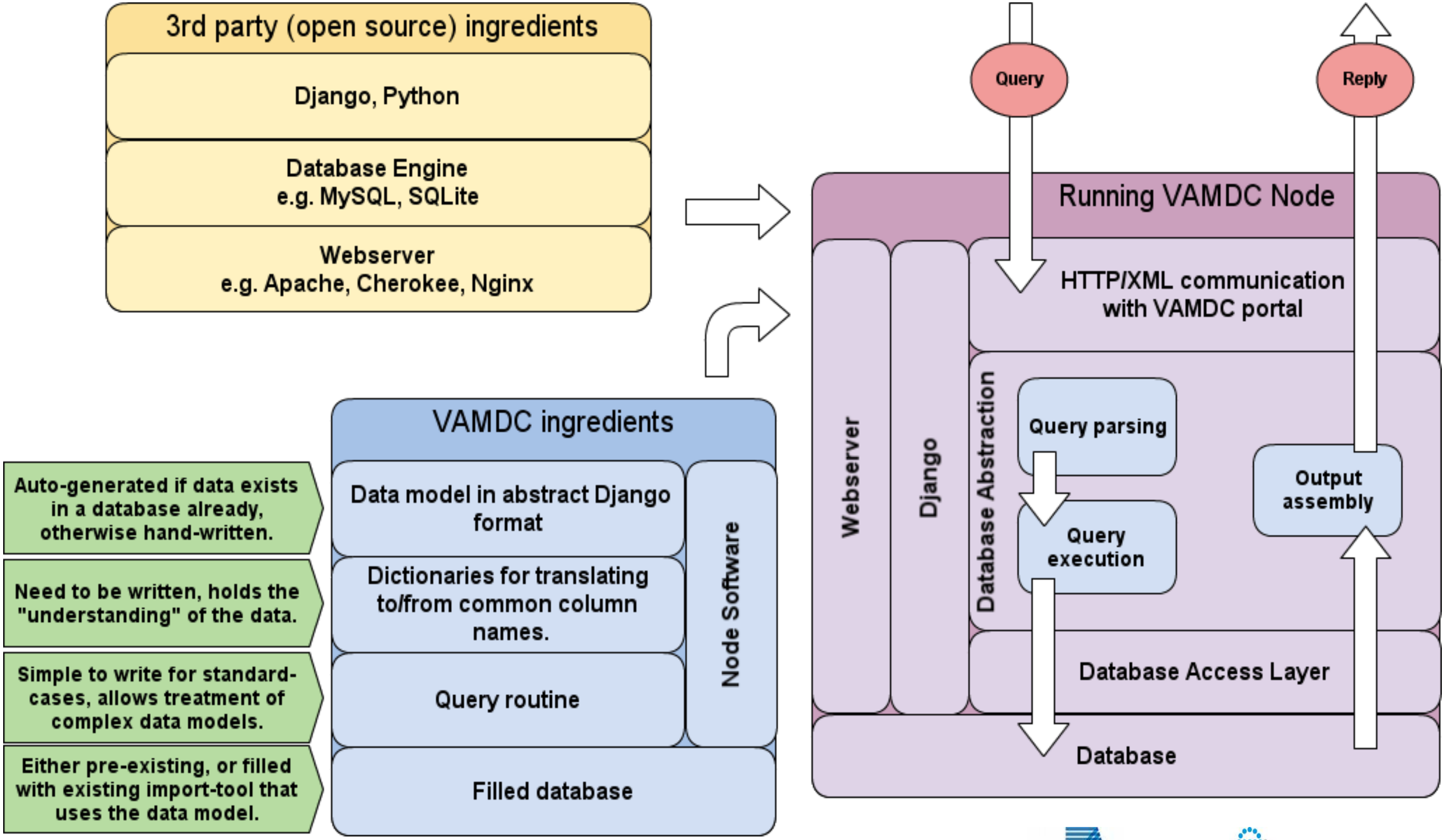
- XSAMS
- Node software
- Registry
- Portal
- Publishing tools
- “Consumer tools”

XSAMS

XML Schema for Atomic, Molecular and Solid Data

- XSAMS is a great idea as it makes the data self-documented.
- XSAMS is even better as it contains (in a sense) large fraction of atomic and molecular physics and since not many people study the subject today it may well be the XSAMS will preserve the knowledge for the future generations.
- XSAMS is a monster and as any XML document is unreadable
- XSAMS is nearly complete for atoms, far from that for molecules and quite impossible to generalize to clusters, solids, chemistry etc.
- End-user hates XSAMS because it is incompatible with their applications

Node software



3rd party (open source) ingredients

Django, Python

Database Engine
e.g. MySQL, SQLite

Webserver
e.g. Apache, Cherokee, Nginx

VAMDC ingredients

Data model in abstract Django format

Dictionaries for translating to/from common column names.

Query routine

Node Software

Filled database

Auto-generated if data exists in a database already, otherwise hand-written.

Need to be written, holds the "understanding" of the data.

Simple to write for standard-cases, allows treatment of complex data models.

Either pre-existing, or filled with existing import-tool that uses the data model.

Node software

- Processes queries translating XSAMS names to local names.
- Returns headers giving counts of available data entries.
- Generates XSAMS on the fly (if requested) complementing data with bibliography.
- Is not portal-specific and as such can be queried directly by automatic and interactive tools (see "Consumer tools").

Dictionaries

- A way to avoid major modifications of individual databases: each data type stored is matched to the corresponding XSAMS name complemented with long a short description.
- Dictionaries are used to parse the query and to generate an XSAMS.
- Dictionaries were produced by the local database groups with the mutual help.
- The list of XSAMS names was compiled by “flattening” the corresponding hierarchy and sometimes shortening it. These together with descriptions are available through a separate tool.

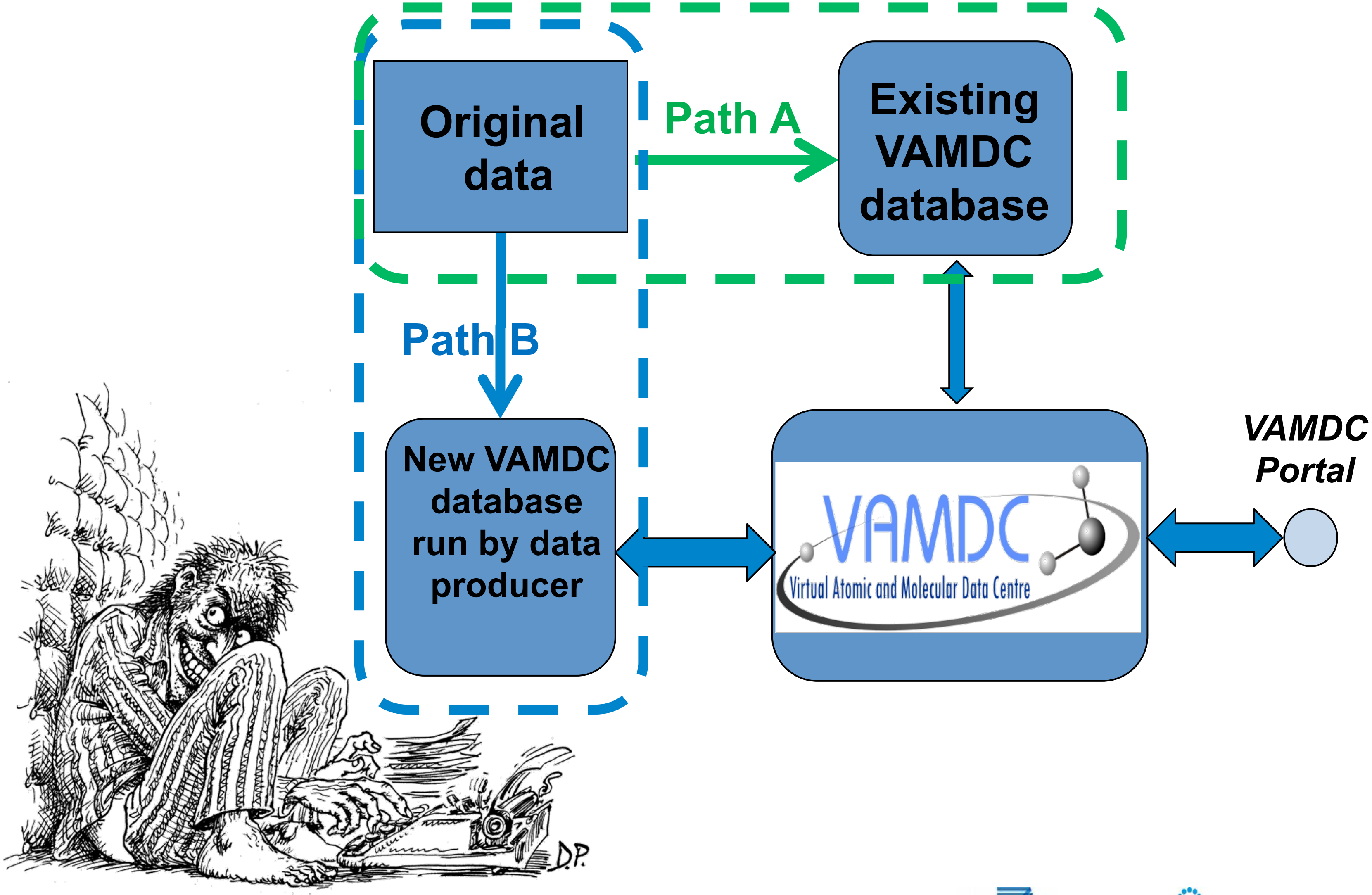
Registry

- Inherited from VO
- Upgraded to regularly query headers of individual nodes and thus give the portal the ability to preselect the nodes
- Another new feature – automatic switching to mirror sites if the main one is down
- New registry entries are added manually – so this needs an administrator

Portal

- Many debates and many versions.
- Division to species and processes.
- Use of InChI codes for the species in addition with other names. The corresponding DB is maintained as a part of the VAMDC infrastructure.
- Portal is an interactive tool that helps constructing a valid query.
- It also determines nodes that have relevant data and sends the query to the selected ones.
- The results can be returned to the user or passed to the Consumer tools.

Publishing tools



Original data

Path A

Existing VAMDC database

Path B

New VAMDC database run by data producer

VAMDC

Virtual Atomic and Molecular Data Centre

VAMDC Portal

Publishing tools

- A simple way to add new data to an existing VAMDC node: if the node contains similar data, just importing it to a DB is all that is needed.
- Data producers can follow the same practice as all the VAMDC nodes and create a new node. We keep all node-specific components of the node software in the repository as examples and we have documentation and tutorials on how to do this.
- Finally, if nothing helps, we can come and do it with you.

Consumer tools

- Once the portal identified nodes with relevant data the user has an option to redirect the XSAMS to the Consumer tools.
- Consumer tools essentially convert XML into a table presenting a projection of the data structure.
- For example, for radiative processes we have a tool that creates tables of energy levels sorted in excitation energy and a table of radiative transitions sorted in wavelengths. The two are cross-linked through energy level labels.

Consumer tools

- We create tools that a generic (e.g. extract the bibliography) or application-oriented creating tables that can be directly used by popular software packages (e.g. XSAMS2SME tool).
- While we have the documentation and examples on how write a consumer tool we are also prepared to provide services in creating such tools on demand.

Lessons learned

- In such projects one must have the flexibility to drop and add partners in the process.
- For the boss one should find the coolest person in the consortium who is willing to do it.
- The workshop format was very efficient both in terms of costs and results.
- Many of the initial ideas were implemented
- A few things did not (like data merging).
- Next are my personal impression of success and mistakes of the VAMDC project.

Success stories

- Node software: works with a variety of commercial and open source relational databases. For an existing DB one should only create the XSAMS dictionary. If no DB exists on must also describe the relations between the data and the node software will generate a new database.
- Dictionaries: solved the issues of interpreting the queries for individual nodes, presenting the extracted results as an XSAMS object and supporting the query creation at the portal.

Success stories

- Portal: efficient, responsive, good-looking. Special advantage is the real-time interaction with the registry to assess the status of the nodes and their relevant data content.
- Registry: recycling of the VO standard with only minor modification, support of the mirror sites.
- XSAMS: quick advance of the standard based on the content of the nodes. Achieving nearly complete description of atoms. Introduction of the case-by-case description for molecules.

Success stories

- Nodes: massive cleaning of inconsistencies, typos and simple errors. Anything from log*gf* truncated from -15 to 15, to major mismatch between the given wavelength and the energies of the levels involved. In some cases, we even managed to provide useful feedback to data producers.
- Nodes: inclusion of quantum numbers and homogeneous description of energy levels.
- Possibility of comparing data from different sources.

Mistakes and failures

- Ambition to return a single dataset to the user was beyond reach.
- The choice to return the data to the user in XSAMS format was was bad. Most of the applications require data in table format. This issue was realized early on and the consumer tools partially alleviated the problem.
- XSAMS data structures are very large and even compressed this imposes restriction on the amount of data that can be requested.

Mistakes and failures

- Some of tools available through the native interfaces of the nodes could not be ported to the VAMDC directly. For example, line strength calculations in HITRAN and Chianti and EXTRACT STELLAR in VALD. These are model-based extraction that requires the parameters of the environment (temperature, pressure, composition) not implemented in the portal and some parameters of the species (e.g. partition functions) not supported by the XSAMS.
- Implementing these will need some further work on standards and software.

Conclusions

- VAMDC was a fun project to work on.
- It already resulted in major improvement of data quality through cleaning errors and forcing data producers and data collectors to be more accurate and systematic.
- Joining VAMDC is relatively easy.
- The mistakes that were made are not fundamental flaws but repairing them will require extension of the standards (XSAMS) and additional work on the portal and the node software.