Virtual Atomic & Molecular Data Centre: technological update

Guy Rixon

EuroVo ICE technical forum

May 2012

VAMDC aims

- Uniform access to A&M data
  - Data formats
  - Web services
- *Like* the VObs, but self-contained
- Serve many areas of science
- Efficiency!
- Preserve provenance, attribution of data

The data

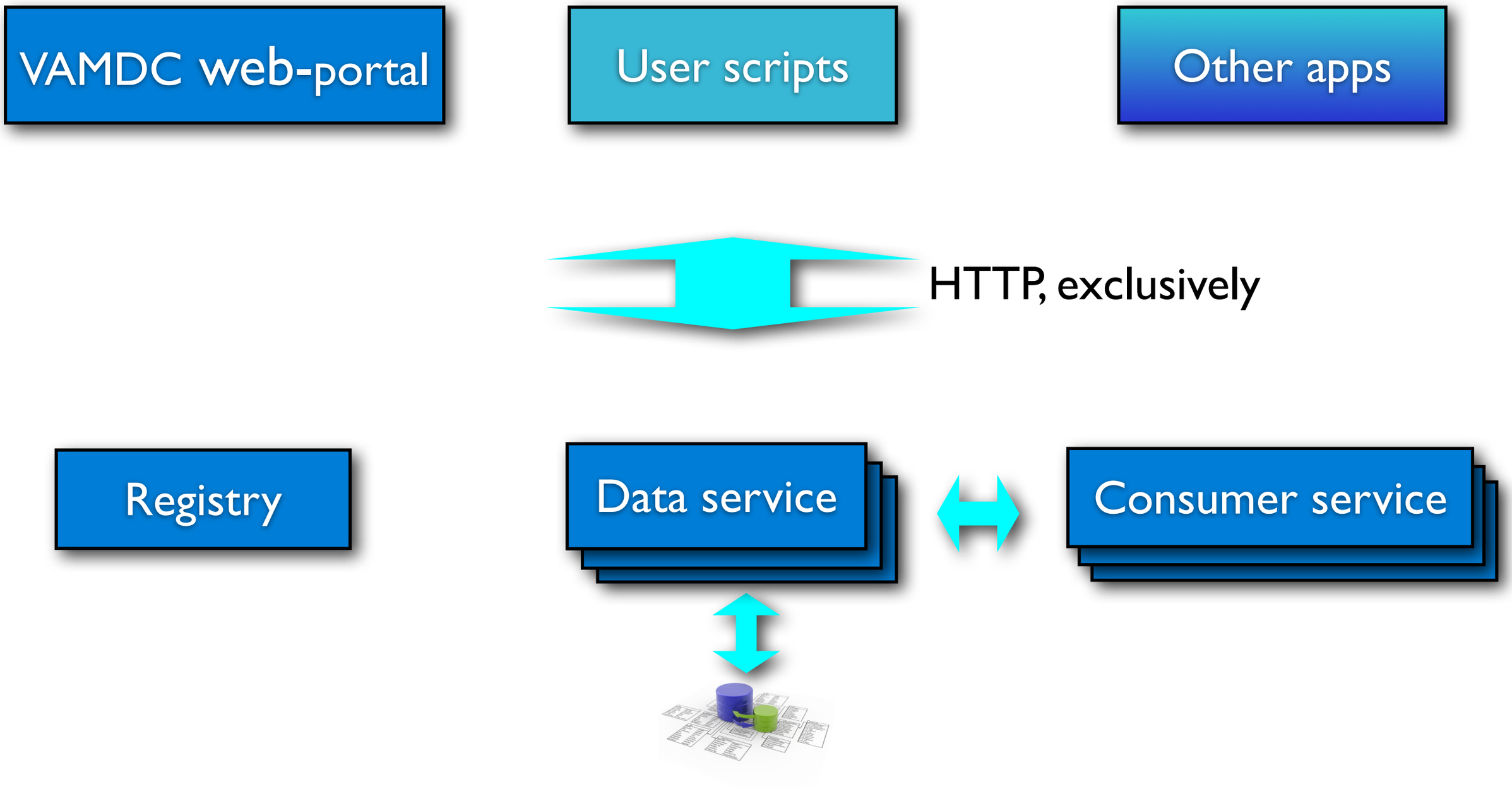
- *Lists* and *tables* of:
  - Atomic/molecular states
  - Transitions between states
  - Lines arising from atomic transitions
- *Not* images; rarely spectra
  - From lab measurement, or from theory
  - Many "small" DBs: MB up to ~10GB

Transitions can be collisional or radiative.

VAMDC as a project

- EU FP7, ~€3M
- 3 years + intro + outro
- Started July 2009, ends December 2012
- 15 paid partners + unpaid collaborators
- ~30 FTE but only ~10 FTE coding
- http://vamdc.eu/

Architecture



Web services and clients, connected with HTTP on public internet.
Two-layer architecture: any client can talk to any service.
Consumer services can be directed to get data from data services: avoids routing data through clients.
One data service per database.
Basically as per VObs.
VAMDC provides services, portal, few clients; community provides remaining clients.

Evolving set of services

VOSpace CEA

UWS

TAP **2009** Grid!

SLAP Registry

etc.!

TAP-XSAMS

TAP CEA

2010 Grid

(SLAP) Registry

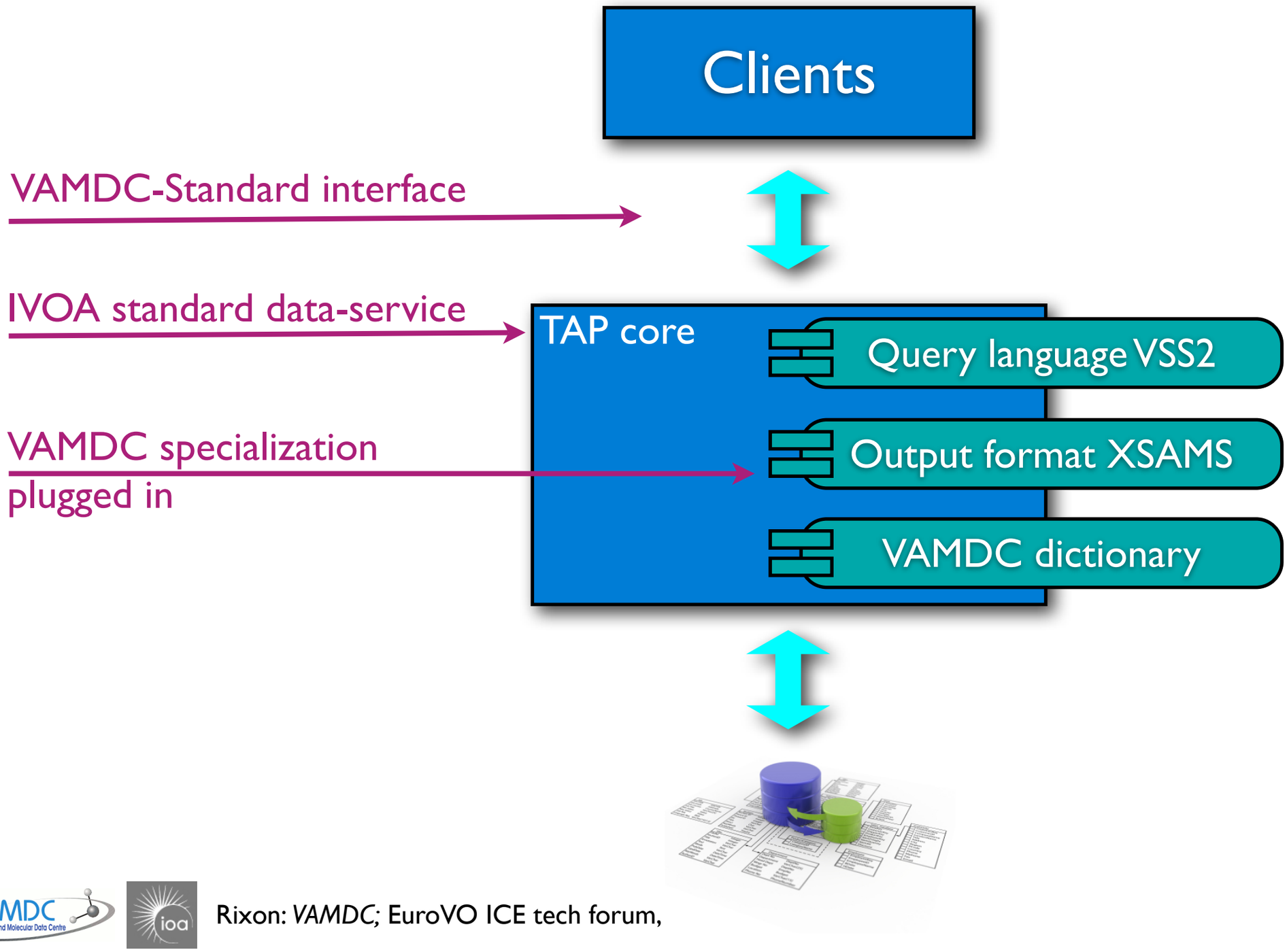
VAMDC-TAP

Registry **2012** (Grid)

XSAMS-consumer

2009: many service types; maximal (planned) use of IVOA standards + gridishness
2010: some IVOA services abandoned as unnecessary; SLAP reduced to single broker-service; grid de-emphasised; advent of specialized TAP-XSAMS protocol
2012: TAP-XSAMS renamed VAMDC-TAP and now sole data-access protocol; all IVOA dropped except registry; grid available (at OPM) but a side-line and not essential to VAMDC use; advent of VAMDC-consumer services

VAMDC-TAP
(né TAP-XSAMS)



VAMDC-TAP was supposed to be a conforming super-set of IVOA TAP. It's diverged and no longer strictly confirms. But it could be brought back to conformance.
There are 22 "VAMDC nodes" providing this protocol, using 3 different implementations

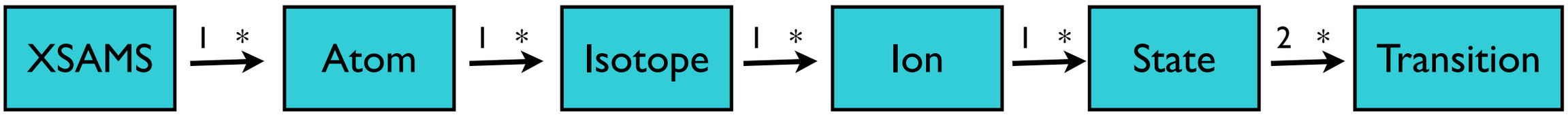
VAMDC-TAP (2)

- Small DBs ⇒synchronous queries only
- Standard pseudo-schema (dictionary) ⇒ no TAP_SCHEMA or VOSI equivalent
- "Try before you buy": HEAD request gets counts of species, states, processes.
- http://vamdc.eu/documents/standards/dataAccessProtocol/index.html

The HEAD requests are crucial to making VAMDC useable interactively: allows iterative refinement of queries. HEAD times are <30s; GET times may be several minutes.

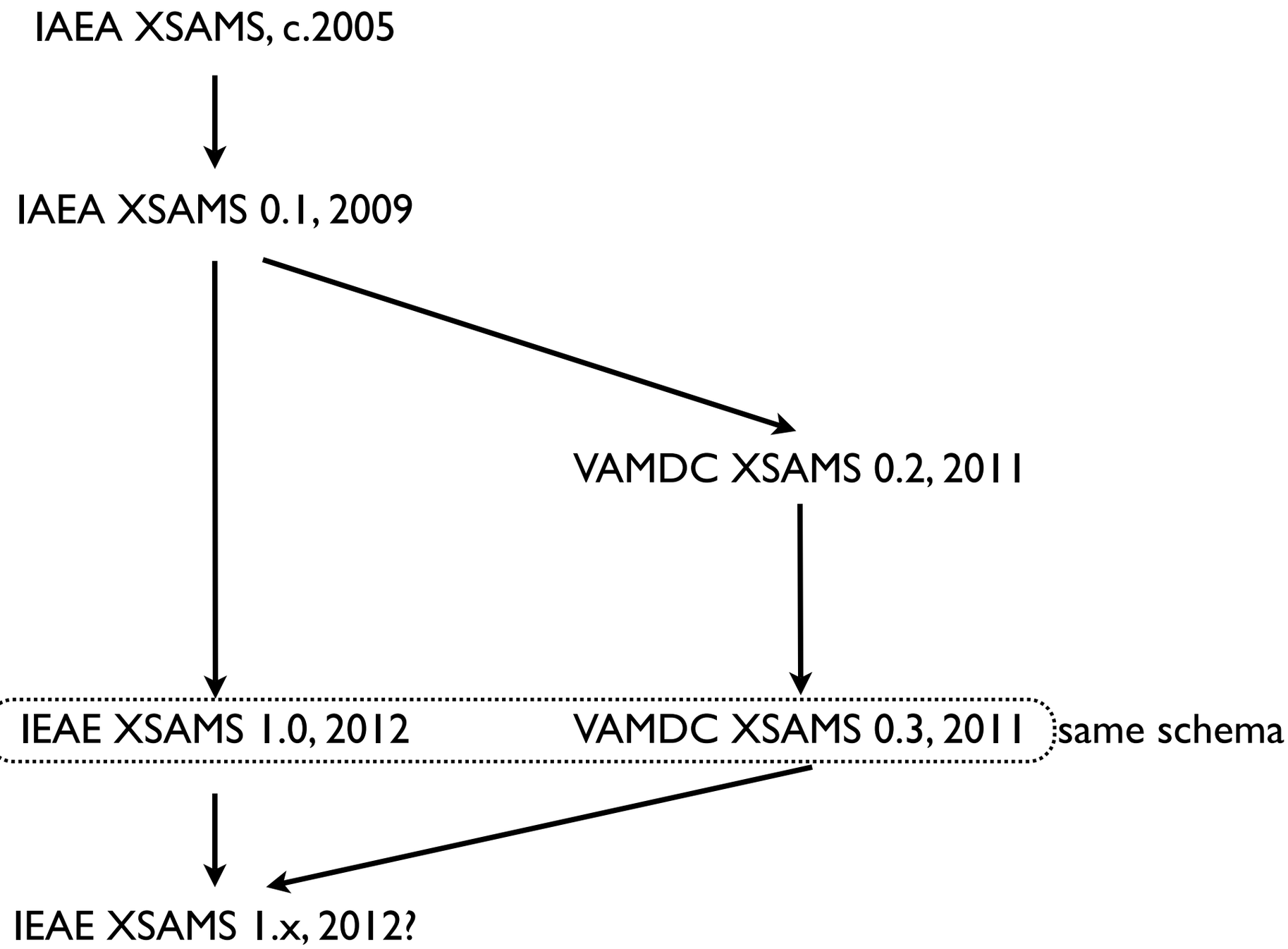
XSAMS (1)

- **X**ML **S**chema for **A**toms, **M**olecules and **S**olids
- Defines both data model and serialization
- Many-to-one relationships all through
- http://vamdc.eu/documents/standards/dataModel/vamdcxsams/index.html



XSAMS is NOT isomorphic with a VOTable, or with any single table.
It COULD be isomorphic with a very-complex schema of tables with interrelations.
This comes from the physics.
Any single-table representation of A&M data simplifies the underlying model.

XSAMS (2)


IAEA XSAMS, c.2005
IAEA XSAMS 0.1, 2009
VAMDC XSAMS 0.2, 2011
IEAE XSAMS 1.0, 2012
VAMDC XSAMS 0.3, 2011
same schema
IEAE XSAMS 1.x, 2012?

VAMDC dictionary

- Standard names for quantities and concepts
- E.g.
  - MoleculeStoichiometricFormula
  - RadTransWavelength{Value, Accuracy, Unit, ...}
- Uses:
  - query language;
  - internal configuration of data-services
- Implied tabular data-model
- http://dictionary.vamdc.eu/

Each dictionary entry has name, description, data-type, and preferred unit.
Entries for observable quantities come in groups {Value, Accuracy, Unit. etc.}
Dictionary terms act as column names in a virtual DB table (one, standard table only)

VSS2

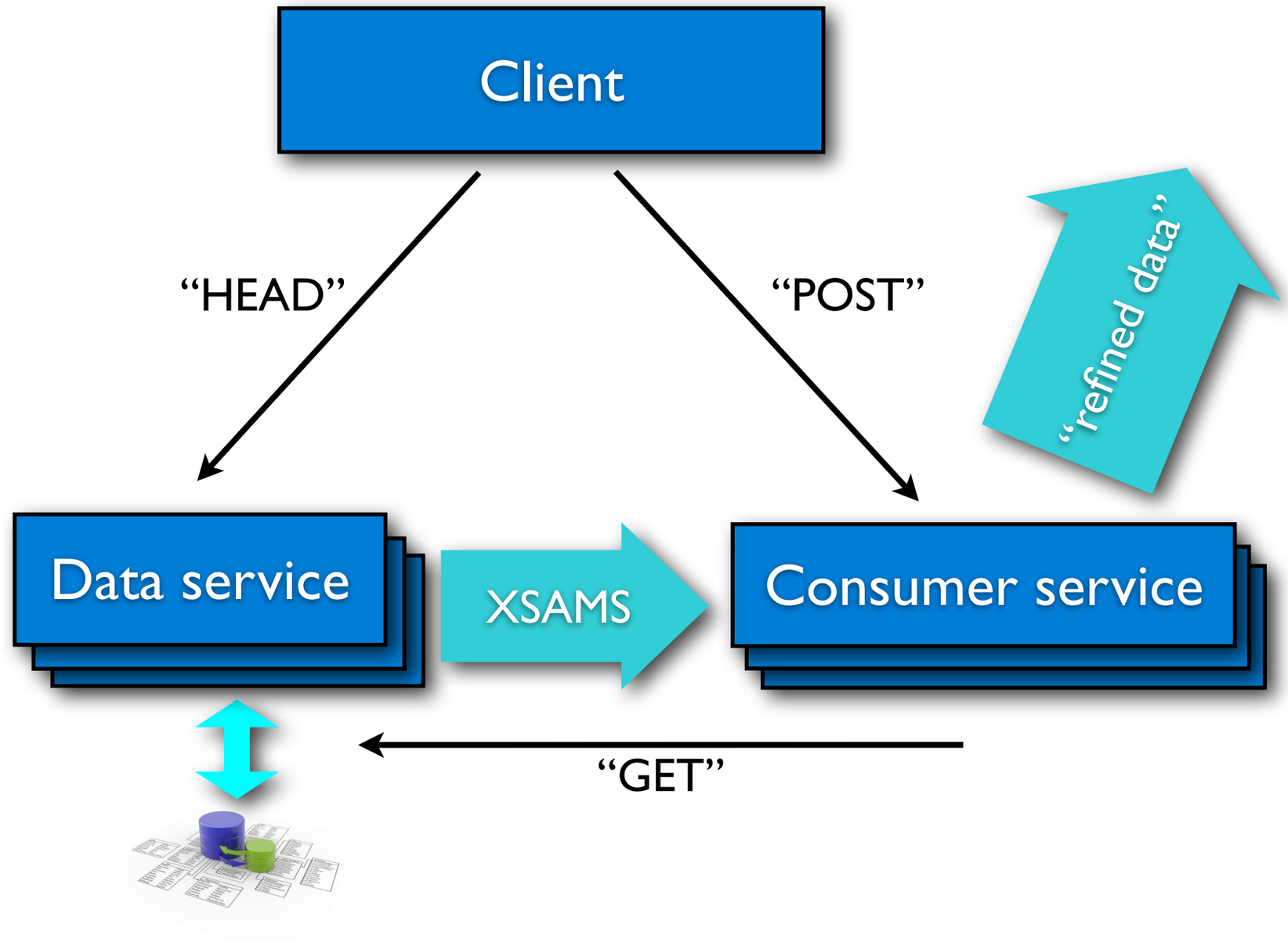
SELECT species WHERE MoleculeStoichiometricFormula='CO'

SELECT ALL WHERE
target.MoleculeStoichiometricFormula='CO'
AND collider.AtomSymbol='He'

- **V**AMDC **S**QL **S**ub-set #**2**
- Operands from VAMDC dictionary
- http://vamdc.eu/documents/standards/queryLanguage/

VSS1 -> VSS2: requestables, e.g. select species instead of select all; prefices, e.g. target, collider
Standard pseudo-schema implied by dictionary => no from clause: single-table view, even when actual DB has multiple tables

XSAMS consumers (1)



Need for filters on XSAMS output: e.g. format conversion, combination of data-sets, visualization.
Implemented as web services (as well as libraries).

XSAMS consumers (2)

- Few so far:
  - Tabular views of XSAMS results
  - Extraction of "Spectroscopy Made Easy" input
  - BibTeX extractor
- http://vamdc.eu/documents/standards/dataConsumerProtocol/index.html

SME is a stellar-atmosphere-modelling code.

Demonstration

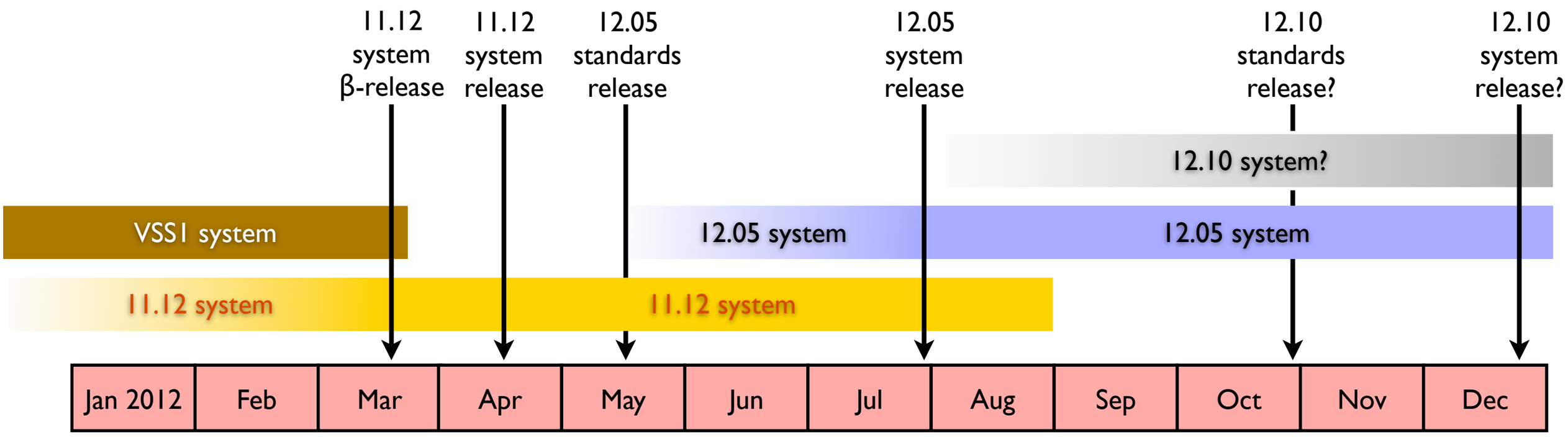
- http://portal.vamdc.eu/
- You can try this system BUT:
  - It's pre-release
  - Please register as beta-tester: email marie-lise.dubernet @ obspm.fr

Public release imminent! Maybe this month!

Reflections: XSAMS

- Adopting XSAMS changes everything:
  - Data model
  - Preferred query-language
  - Service-implementation tech
  - Release management
  - Query model
  - Provenance handling

Reflections: release policy

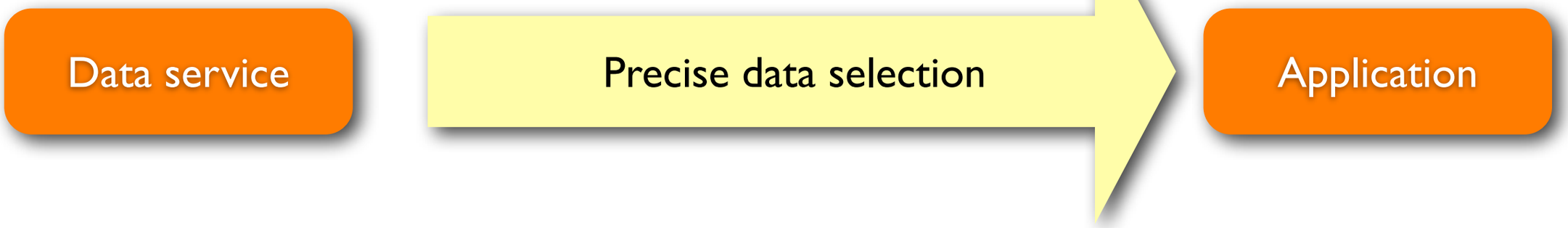


- xx.yy numbers refer to version of standards
- 11.12 system is the one demo'd at PM3 (Vienna, March 2012)
- "12.05" might be 12.04 or 12.06

VAMDC has versions of complete sets of VAMDC standards.
Therefore, it has versions of matching systems.
This is necessary because the XSAMS namespace URI changes and breaks everything.
"System" = registry + portal + compatible data services + compatible consumer services
At any one time, we have, at least, a released, stable system and a development system.
This is possible because VAMDC is a small project and well coordinated (and because the participants are diligent and cooperative).
In future, we expect less-frequent changes to standards - perhaps once per year, perhaps less often still.

Reflections: query model

Traditional IVOA approach:



Emerging VAMDC approach:

VAMDC is tending towards a system where a coarse cut from archives is transferred and cached as XSAMS and then multiple “queries” are run on that file to feed an application. I.e., parsing the XSAMS involves a selection process, something like a query, possibly mediated by XPath or XQuery.
The fine “query” may typically be done by a library rather than a service.
My opinion: this is a good thing, because the XSAMS intermediate-product retains the metadata and provenance.

Reflections: provenance

- XSAMS is (maximally?) self-describing
- Every point in the data model can have provenance and attribution.
- Very good for data published in support of papers.
- Also, can cite the original data providers
- This is VAMDC policy (part of our EU remit)
- Trial being arranged with IoPP to "enforce" data citations in papers.

If you want to publish the (A&M) data supporting a paper, XSAMS is a good way to do it – the XSAMS file needs little supporting information and is inherently queriable without the need for a data service. If the XSAMS can be lodged in a proper repository (e.g. institutional library), the job is done.
The XSAMS can also carry the attribution of data origins, both for the VAMDC node and for the data fed into that node (typically a node aggregates results from many workers). It is a specific goal of VAMDC to cite the original producers of data in papers written by users of VAMDC; this needs the bibliographic content of XSAMS and some editorial pressure by the journals.