Tools for manipulating/mining data

Ljerka Nenadović
Université Pièrre et Marie Curie LPMAA CNRS

Cycle One Project Meeting, Milton Keynes, UK, 19-22 April 2010

Goal

- demonstrate extraction and manipulation of data from species represented in XSAMS
- perform energy level matching between data representing the same species based on quantum number values

Obstacles

- Unique identification of species

Solution:

Two steps:

1. Request only summary information for all species of a given molecule
2. Ask the user to specify the species based on the summary information before retrieving complete XSAMS for the selected species.

Summary information

- What are the fields mandatory for describing a single species?
  - How do we distinguish different isotopes in XSAMS?
- What about different measurements/calculations representing the same species?
  - Unique identification of sources

Need to map mandatory fields to fields in XSAMS

Initial summary information fields:

1. molecule
2. stoichiometric formula
3. textual description
4. reference/source
5. database

Obstacles

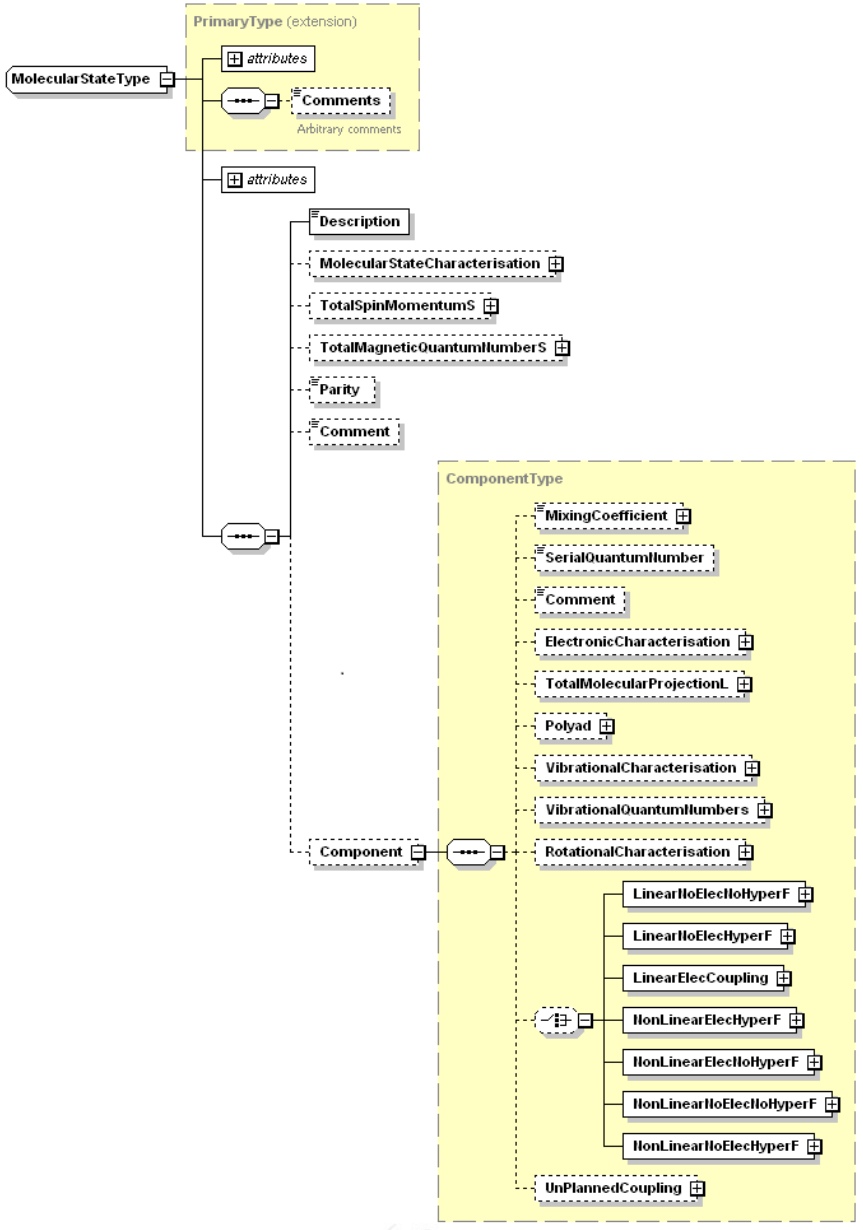
- deeply nested XSAMS structure

  Solution:

  Propose that XSAMS be simplified by flattening the structure.

  - Changes discussed at the last XSAMS meeting in Japan

Flatter XSAMS schema



Problem: existing implementations use "old" XSAMS

Solution: develop a method for parsing xml files and outputting new xml files with a flatter XSAMS structure

Obstacles

- sorting of energy levels

Solution:

The best approach would be to sort as part of the database query before turning the result into XSAMS, however since I was dealing with flat files, as well as SQL databases, I had to develop an algorithm for sorting molecular states. This was done based on the energy value.

- **Consideration**: units
  - Need to develop a list of acceptable units and a standard list of routines for transferring values from one set of units to another
- **Consideration**: different origin
- **Consideration**: missing levels

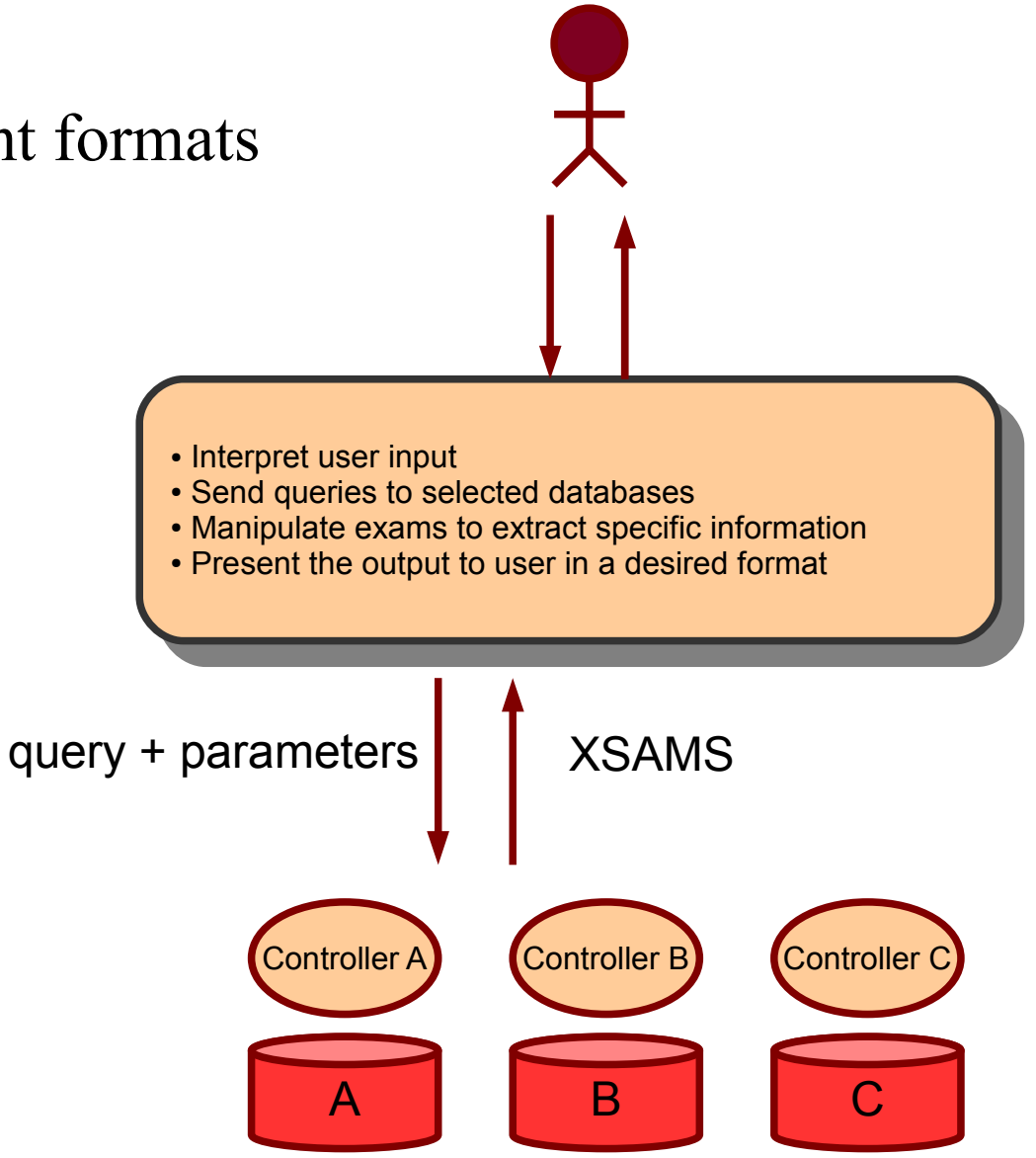
Obstacles

- multiple databases in different formats

Solution:

Create a separate controller

for each database



Extraction of data from XSAMS

- Energy level
- Energy value
- List of quantum numbers

Rotational energy levels of ortho-H2O (cm^{-1})

Kyro (1981)

level	energy	N	v1	Ka	Kc	tau
1	23.7943	1		-1	0	1
2	42.3717	1		1	1	0
3	79.4963	2		-1	1	2

Matching

- Molecular states are sorted
- Quantum numbers are extracted and compared

Rotational energy levels of ortho-H2O(cm^{-1})

Kyro (1981) vs Faure et al, A&A, 472, 1029 (2007)

level	energy 1	energy 2	N	v1	Ka	Kc	tau
1	23.794300	23.794360	1		-1	0	1
2	42.371700	42.371740	1		1	1	0
3	79.496300	79.496390	2		-1	1	2

Further considerations

- Expected quantum numbers for specific species
- null vs 0
  - null = no data available
  - 0 – valid field value

which one is correct ?

Rotational energy levels of para-H2O (cm^{-1})

Faure et al MNRAS 347, 323 (2004)
vs Faure et al, A&A, 492, 257 (2008)

l	energy 1	energy 2	N(1)	N(2)	v1(1)	v1(2)	tau(1)	tau(2)	Ka(1)	Ka(2)	Kc(2)	Kc(2)
1	0.0	0.0	0	0		0	0	0	0	0	0	0
2	53.4	37.13	1	1		0	0	0	1	1	1	1
3	100.9	70.09	2	2		0	-2	-2	0	0	2	2