Q5cost

A common data format for program interchange in the

Quantum Chemistry domain

Elda Rossi
CINECA – Bologna (Italy)

The problem

- The problem was driven by a common practice in the group:
  a scientific problem needs several programs to be solved:
  - specific in-house codes written by some of the partners (FCI, CASDI, NEVPT, PROP, TouChain).
  - open source or commercial programs used for producing the standard quantities (COLUMBUS, DALTON, MOLCAS, MOLPRO, GAMESS, …)

- None of them shares the same data format
- Tricky procedure
  - Get all the programs
  - Install them on the local computer
  - **Translate the data**

Elda.Rossi@CINECA.it - Bologna (Italy)

The solution: a computing/collaborative Grid

- Each program remains “at home”
- Each program maintains its own proprietary data format
- Each program must understand a common data format for interchange
- A set of “translation utilities” takes care of translating data from/to the proprietary formats and the common format
- A sort of “grid environment” makes programs communicate and exchange data through a single interface

Elda.Rossi@CINECA.it - Bologna (Italy)

The final vision

- Resources (mainly applications)
  - geographically distributed
  - on heterogeneous platforms

  (resource GRID !!!)
- A central (web) interface for accessing them
- A way to define the flowchart and to run it **(workflow!!!)**
- Dynamic control over the execution (steering)
- Data are collected in a (virtually) central repository

Elda.Rossi@CINECA.it - Bologna (Italy)

The first problem: Data Format

- One of the main challenges is a lack of community standards for data representation.

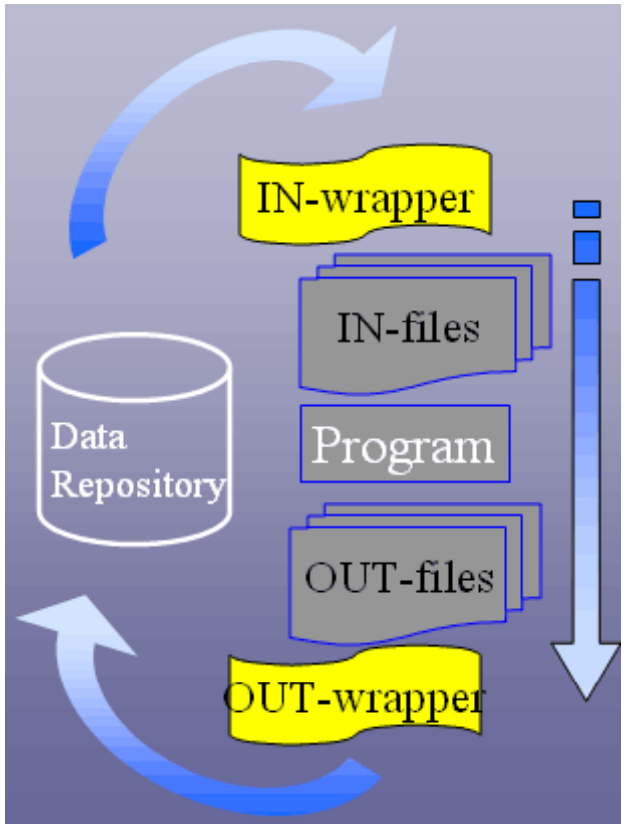
- Two main strategies:
  - To develop translators that map one format to another. This requires n**2 translators, additional data formats requires 2n more translators (theoretically be very unscalable)
  - Support the development a community standard for the field and develop translators towards the common formats.

- Other – more advanced – alternatives exist
  - if all the data formats being used are represented as XML, third party techniques such as XSLT can be used to perform the translations.
  - If not, metadata languages such as Data Format Description Language (DFDL) can be used to describe the data.

Elda.Rossi@CINECA.it - Bologna (Italy)

The Common Data Format

- Define a data format for interoperability
- What do we need
  - Complete
  - Flexible
  - Good performance
  - Near to chemists
  - External
- How to use it
  - Data repository
  - Input/output wrappers for data conversion
- A format for **interchange**, not to be used as an internal format
- We are interested in **functionality** (that has to be general and complete). **Performance** and efficiency, although important, are not the main focus.



Elda.Rossi@CINECA.it - Bologna (Italy)

The data model

- We identified **two different kinds** of information
  - **small data** quantities, mainly ASCII coded, like atom labels, geometry, symmetry, basis sets and so on
  - **large datasets**, normally binary, like integrals and expansion coefficients.

Elda.Rossi@CINECA.it - Bologna (Italy)

Small/large data

- **Small data**: several initiatives, all based on XML (human readability and machine comprehension).
- We have adopted the same approach ➔ **QC-ML** (Quantum Chemistry Markup Language)
- For processing QC-ML files we need a specific FORTRAN library ➔ **f77xml**

- **Large datasets**: XML is not convenient, mainly due to its verbosity.
- **HDF5** was adopted ➔ **Q5cost**
- For processing Q5cost files we need a specific FORTRAN library ➔ **Q5cost**

Elda.Rossi@CINECA.it - Bologna (Italy)

Small data

Base facts

Symmetry: the symmetry of the system in terms of group name and other symmetry data;

Geometry: the atomic composition of the system and its cartesian coordinates;

Basis: the basis set information, either given by name or fully defined.

Derived facts

Energies
Properties
Integrals
Coefficients
....,.

Large !!

- All these data are rather "small" and can be effectively described using a mark-up language for enhancing readability and standardisation.
- A hierarchical scheme was designed and described based on XML
- we called it QC-ML (Quantum Chemistry Mark-up Language).

Elda.Rossi@CINECA.it - Bologna (Italy)

QC-ML: an XML format for QC

Base facts

Derived facts

```
<molecule nElectrons charge
  spinMultiplicity spaceSymmetry>
   <symmetry ... />
   <geometry ... />
   <basis    ... />
</molecule>


<computedData>
   <energy … />
   <property … />
   <Q5file address URL/>
</computedData>
```


Elda.Rossi@CINECA.it - Bologna (Italy)

Large bin data

- We looked for a suitable technology that can merge
  - portability,
  - efficiency,
  - FORTRAN binding,
  - data compression, and
  - easy access to information.
  - Usability is also important but not critical

HDF5 was considered the right technology

Elda.Rossi@CINECA.it - Bologna (Italy)

What is HDF5 ?

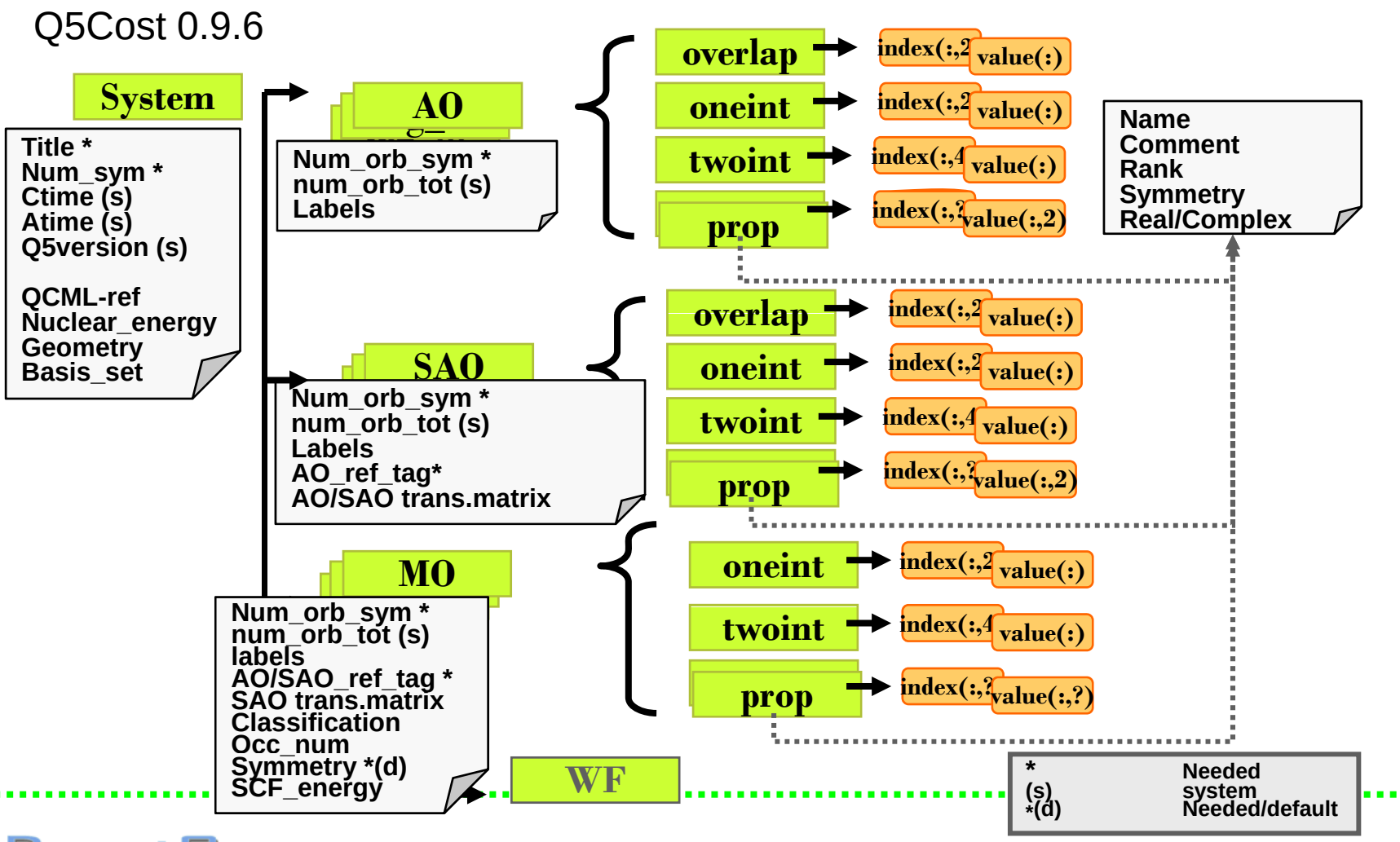
- HDF5 is a Format and software for scientific data produced by NCSA/University of Illinois
- Support any kind of data for digital storage regardless of origin and size
- Stores data in a highly organized and hierarchical format
- High efficient chunked I/O
- Allows inclusion of metadata (attributes)
- Platform independent file format
- Widely used in scientific or visualization codes

http://www.hdf-group.org

Elda.Rossi@CINECA.it - Bologna (Italy)

The hierarchical structure

All the data objects are related within a **hierarchical structure** and **logical containment relationship**



Elda.Rossi@CINECA.it - Bologna (Italy)

Tools: library

```
Q5Cost_MOOneInt_create
    (file_id, error, [order], [user_tag] )
```
Creates a new container for MO one electron

```
Q5Cost_MOOneInt_append
    (file_id, index, value, howmany, error, [user_tag] )
```
Appends MO one electron integral values to the file

```
Q5Cost_MOOneInt_read
    (file_id, offset, howmany, index, value, error, [user_tag] )
```
Reads MO one electron integral values to the file

```
Q5Cost_MOOneInt_clear
    ( file_id, error, user_tag )
```
Appends MO one electron integral values to the file

Chemists prefer Fortran:

- **F90**
  the general HDF5 library distributed by NCSA, we did the specific Q5cost interface
- Other “bindings” produced authomatically by a python script: C, C++, F77

Elda.Rossi@CINECA.it - Bologna (Italy)

Tools:

- **Q5dump:** allows you to see the content of a Q5 file
- **Wrappers**: a program, specifically designed for each QC code in the chain, capable of retrieving information from, and writing information to, the file in accordance with the defined syntax. It ises the Q5cost library.
  - **input wrapper** reads data from the repository file and converts them into the QC code specific input,
  - **output wrapper** reads data from the QC code specific output and adds them to the repository file.

Elda.Rossi@CINECA.it - Bologna (Italy)

Wrappers

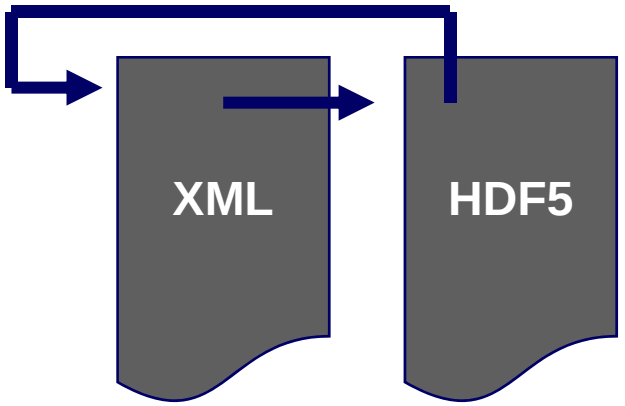
- FullCI: embedded
- TouChain: wrapper via MolCOST, embedded in progress
- MRCC: in progress

- Dalton: wrapper
- GamessUS: embedded in progress
- MolCAS: wrapper
- Columbus: wrapper
- OpenBabel: in progress
- Molekel (via OpenBabel): in progress

Elda.Rossi@CINECA.it - Bologna (Italy)

Conclusion and future work

- I discussed our solution for code integration (in QC context)
- the data model consists in two parts
  - Small data ➔ XML
  - Large data ➔ HDF5



- A hierarchical structure of both of them has been proposed
- Some sw tools (Fortran libraries) are available designed on the data model

- Enhancement of the data model
- Collaboration towards a common data model
- Put everything on the “grid”

Elda.Rossi@CINECA.it - Bologna (Italy)

Acknowledgment

COST D37 project:
Grid Computing in Chemistry- GRIDCHEM

- DeciQ Working Group:
  Code interoperability in Computational Quantum Chemistry

– **Elda Rossi**, Andrew Emerson – CINECA
– Gian Luigi Bendazzoli, Antonio Monari – Univeristà di Bologna
– Renzo Cimiraglia, Celestino Angeli, Chiara Pastore - Università di Ferrara
– Daniel Maynau, Stefano Evangelisti, Anthony Scemama - IRSAMC – Toulouse
Vallet Valerie, JeanPierre Flament (University of Lille)
– José Sanchez-Marin - Universitat de Valencia
– Peter Szalay, Attila Tajiti - Eötvös Loránd University, Budapest
– Kállay Mihaly (Budapest University of Technology and Economics)
– Baldridge Kim (University of Zürich)
– Kenneth Ruud (University of Tromsø)
– Stefano Borini (now at DTU – Denmark)

Elda.Rossi@CINECA.it - Bologna (Italy)