Data representation and code interoperability for computational materials physics and chemistry

Alberto Garcia
Universidad del Pais Vasco, Bilbao (Spain)

Phil Couch
CCLRC-Daresbury Laboratory (United Kingdom)

Thomas Schulthess
Oak Ridge National Laboratory (United States)

CECAM (Lyon, France), 19, 20, and 21 April 2006

Summary

The computational approach to the study of matter has been hugely successful, but this success has brought some practical problems. On the one hand, the inmense amounts of data generated need to be analyzed and/or archived and catalogued for later retrieval or distribution. On the other, it is sometimes desirable to study the same system with various codes which implement different features or levels of approximation (e.g. a tight-binding calculation followed by an ab-initio one). Currently, the input/output formats of the vast majority of codes are completely different, even if they refer to magnitudes (i.e., atomic coordinates, charge densities) which are perfectly well characterized from the physical point of view.
This workshop was devoted to the use of data representation tools (basically XML and related technologies) in atomistic simulations, with an emphasis on the needs of the electronic structure community in regard to interoperability of codes and ease of post-processing of results.

This event was made possible thanks to partial support from:

- **CECAM**
- **The Psik network**
- **The European Science Foundation** (ESF) under the EUROCORES Programme **EuroMinScI**, through contract No. ERAS-CT-2003-980409 of the European Commission, DG Research, FP6.

Workshop Conclusions

1. The workshop highlighted several key motivations for the development of a data management framework and these are summarised below.

   - *Interoperability*
     Increasingly, addressing many scientific problems requires the use of more than one application used in close co-operation. For example, often, any attempt to bridge length and time scales requires a hybrid of classical and quantum mechanical methods or the application of various levels of theory. Such approaches could be implemented as a workflow, with information being exchanged between various simulation codes. The control of such complexity mandates a coherent approach to the management of data.
   - *Visualisation and analysis*
     Considerable effort is invested in the development of tools that can be used to analyse and visualise simulation results. Unfortunately, these tools are often restricted to use with specific applications.
   - *Data longevity*
     Currently, it is often quicker to duplicate effort regenerating simulation results than to locate and retrieve existing results from a repository; data longevity requires serious consideration.
   - *Data volume*
     Data are generated at ever increasing rates and methods for dealing with large volumes of data from both experiment and simulation are required.
   - *Efficient use of future computer resources*
     The present computing paradigms don't scale well. The development of data organisation strategies will be key for efficient use of the next generation of machines.

2. A number of groups have started their own efforts toward structured data handling. In the process, several useful tools have appeared. This workshop has served in part as a useful showcase of those efforts and developments. Most work has been done with XML, although there have been important advances in the handling of large binary data sets in portable formats (netCDF, HDF5). Among the XML efforts, the chemical markup language (CML) figures prominently in several projects, while other groups have designed their own data markups.

3. A global standard for data formats in chemistry and materials science is not realistic or practical at present. On the other hand, convergence in narrower domains, and the initial adoption of data representation concepts and tools by most codes might be within reach in the short to medium term.

4. An ontology-mediated approach (that involves the specification of relevant concepts within a given domain) could provide an initial alternative to defining a common global representation for data. Such an approach will allow interoperability between communities that work with different data models and that use different technologies (i.e. XML, relational databases, netCDF/ HDF5).

5. Once well-defined ontologies and data structures with manageable scope are created, a mixed process of federation, nucleation around proven ideas, and pruning of redundant material could potentially result in a successful global schema.

Workshop Recommendations

1. Encourage code developers to look into existing tools for data representation, even if only at the single-code level. The creation of some form of structured data is already a significant step for interoperability, as automatic transformation is in principle possible.

2. Form special-interest groups associated to specific domains for which potential global schema for data structures could be drafted. During the workshop the following domains and people involved were identified:

   - (Molecular) orbitals and basis sets: Knowles, Baldridge, Rossi, Sherwood, Fawzi, Kim, Garcia.
   - Monte-Carlo data: Troyer, Kim,Schulthess.
   - Large binary data (in quantum chemistry and electronic structure): Rossi, Gygi, Fawzi, Garcia.
   - LMTO data sets: Schulthess, Temmerman.

3. Create an infrastructure to serve as follow-up to the workshop and as the basis for future collaboration. The workshop website at www.cecam.fr was used initially to hold the actual workshop material, but immediately afterwards a new Twiki site (www.datarepresentation.org, maintained by P. Sherwood and P. Couch) was created as a more focused and longer-lasting resource to hold links and discuss best practices.

4. Assign initial follow-up tasks

   - Develop a tutorial on dictionaries and schemas: Murray-Rust.
   - Organize an access-grid follow-up meeting a few months after the workshop: Sherwood.
   - Create a table in which input data fields and their use in various condensed-matter electronic structure methods are analyzed: Schulthess, Gonze, Gygi.

5. In the longer term, the community should look into ways to maintain the momentum to develop better data representation frameworks and tools. It seems that there is a good enough case to justify one or several funding applications. There should be a follow-up workshop in about 2 years.

Workshop Impact on the EuroMinScI Program

The workshop was quite successful as a showcase of current approaches to the problem of data representation. There was a very important presence of the e-Minerals group in Cambridge and a contribution by Dan Wilson, who is currently involved in the HydroMin project as part of Björn Winkler's group. The techniques currently used by both groups hinge on making the output of ab-initio and other simulation codes CML-aware, in such a way that the codes can be made part of a complex workflow, including downstream visualization. The presentations had quite an impact on the attendees. We can expect that, in the short term, most codes used by the minerals community will be so instrumented.
Perhaps the most important impact of the workshop on the EuroMinScI program will begin to be felt in the medium term, when the efforts toward a coherent representation of the operational parameters of the codes (i.e., basis sets and pseudopotentials used, etc) reach the production stage. It will be then much simpler to compare the results of calculations at different levels of quality, and to perform cascades of simulations (going from the simpler to the more sophisticated).
Using the tools presented in the workshop, which will be developed as detailed in the conclusions and recommendations, the EuroMinScI and the minerals community in general could perhaps develop a field-specific ontology that could serve as a framework for data dissemination and comparison of theory and experiment. This is a very ambitious development, which falls outside the scope of the workshop itself.

Program

Day 1: April 19 2006

Session : 1 Introduction and Showcase-1
09:00 to 09:10 : Welcome
Alberto Garcia

09:10 to 09:40 : Presentation
Information Exchange in Computational Chemistry and Materials Physics - An Overview
Phil Couch

09:40 to 10:10 : Presentation
Overview of CML
Peter Murray-Rust

10:10 to 10:40 : Presentation
MaterialsGrid and CML
Daniel Wilson

10:40 to 11:10 : Presentation
XML I/O Subsystem for Computational Nanoscience
Thomas Schulthess

11:10 to 11:30 : Coffee Break

Session : 2 Showcase-2
11:30 to 12:00 : Presentation
Grid Enabled Molecular Science Through Online Networked Environments
Kim Baldridge

12:00 to 12:30 : Presentation
Automatically extracting metadata from CML for combinatorial simulations on the grid
Richard Bruin

12:30 to 13:00 : Presentation
On the structuring of the computational chemistry Virtual Organization COMPCHEM
Osvaldo Gervasi

13:00 to 14:30 : Lunch Break

Session : 3 Ongoing efforts in data formats
14:30 to 15:00 : Presentation
Specification of file formats for NANOQUANTA
Xavier Gonze

15:00 to 15:30 : Presentation
A proposal for a unified norm-conserving pseudopotential format
Javier Junquera

15:30 to 16:00 : Presentation
Data file exchange in Quantum-ESPRESSO
Paolo Giannozzi

16:00 to 16:30 : Presentation
Data representation for biomolecular systems: technical and political challenges
Konrad Hinsen

16:30 to 17:00 : Coffee Break

Session : 4 Demonstration-1
17:00 to 17:30 : Presentation
Demonstration of XML tools: xmlf90
Jon Wakelin

17:30 to 18:00 : Presentation
XSLT transforms
Toby White

Day 2: April 20 2006

Session : 1 Electronic Structure Codes
09:00 to 09:30 : Presentation
XML Schema Design for First-Principles Molecular Dynamics
Francois Gygi

09:30 to 10:00 : Presentation
Object representations for quantum simulations
Jeongnim Kim

10:00 to 10:30 : Presentation
The ALPS project
Matthias Troyer

10:30 to 11:00 : Presentation
Unified XML I/O approach in DFT and model codes for real materials: from crystal structure to magnetic susceptibility
Anton Kozhevnikov

11:00 to 11:30 : Coffee Break

Session : 2 Quantum Chemistry Codes
11:30 to 12:00 : Presentation
Data representation in the Molpro quantum chemistry package
Peter Knowles

12:00 to 12:30 : Presentation

A common format for Quantum Chemistry
Elda Rossi

12:30 to 13:00 : Presentation
Building an Infrastructure for Quantum Chemistry: Data Sharing and Graphical User Interfaces
Paul Sherwood

13:00 to 13:30 : Lunch Break

Session : 3 Discussion session 1
14:30 to 15:00 : Presentation
Material leading to discussion: Data markup hierarchies and interoperation
Peter Murray-Rust

15:00 to 16:30 : Discussion

16:30 to 17:00 : Coffee Break

Session : 4 Demonstration-2
17:00 to 18:00 : Presentation
An Introduction to the Use of XMLDIR and QuickSchema
Michael Summers

Session : 5 Social dinner
Day 3: April 21 2006

Session : 1 Other Tools
09:00 to 09:30 : Presentation
XML and the Vienna Ab-initio Simulation Package (VASP)
Orest Dubay

09:30 to 10:00 : Presentation
Information Exchange in Computational Chemistry and Materials Physics - AgentX
Phil Couch

10:00 to 10:30 : Presentation
Evolutionary Construction of Higher-level Constraints over an XML Language; Formalizing Informal Semantics
Toby White

10:30 to 11:00 : Presentation
CCPN - automatic code generation from UML data models
Rasmus Fogh

11:00 to 11:30 : Coffee Break

Session : 2 Other issues
11:30 to 12:00 : Presentation

Handling of large datasets in Quantum Chemistry
Antonio Monari

12:00 to 12:30 : Presentation
Input handling in CP2K
Fawzi Roberto Mohamed

12:30 to 13:00 : Discussion

13:00 to 14:30 : Lunch Break

Session : 3 Discussion-2
14:30 to 16:30 : Discussion

16:30 to 17:00 : Coffee Break

Session : 4 Concluding remarks and hands-on session
17:00 to 18:00 : Discussion

Presentation List

Information Exchange in Computational Chemistry and Materials Physics - An Overview
Phil Couch
CCLRC-Daresbury Laboratory, United Kingdom
Abstract
A lack of information management in computational chemistry and materials physics is hindering progress in a number of important areas. Increasingly, addressing scientific problems requires a mixture of simulations codes used in close cooperation. As a result, considerable effort has been applied to the development of general tools that can be used to specify and execute complex computational workflows. Unfortunately, it is difficult to realise the full potential of these tools due to a lack of data standards. It is simply not possible to exchange information seamlessly and in an automated fashion between workflow components; format convertors and wrappers must be developed prior to execution. Such information exchange between simulation codes and tools that can be used to visualise and analyse data would also be an important development; common tools could be created to work with a range of simulation codes. The XML standards provide a good starting point for the development of our own data standards; they are mature and generally well adopted. This presentation details the requirements of a common XML data model and looks at how different data model designs meet these. Several common approaches for manipulating XML data from simulation codes are emerging; many building upon others. A summary of the different approaches with their corresponding strengths and weaknesses will be presented.

Overview of CML
Peter Murray-Rust
Unilever Ctr for Mol. Informatics, U. of Cambridge, United Kingdom
Abstract

- CML is now a robust technology which can support a large variety of concepts in computational chemistry and physics.
- It is designed as a core architecture with controlled XML syntax but high extensible through the use of dictionaries.
- These dictionaries are closely linked to individual codes which allows developers to work independently. Many dictionaries can be used simultaneously and this allows a model where dictionaries can later be merged and refactored to represent consensus in the community.
- A core dictionary for scientific units has been developed and this allows developers to translate units between different codes.
- A recent development is that CML documents (and therefore programs) can be validated through the JUMBO software. This has been done for output from CASTEP, SIESTA and DL_POLY.
- Developments in XML-CML include: (a) archiving of computational results in institutional repositories (b) engagement with STM publishers about direct publication of results (c) high-throughput analysis of computational chemistry results. (d) response to community demands - e.g. balancing markup vs speed, descriptions of extended objects, etc.

MaterialsGrid and CML
Daniel Wilson
Univ. of Frankfurt, Germany

Coauthors: Andrew Walkingshaw, Bjorn Winkler, Martin Dove, Peter Murray-Rust
Abstract
MaterialsGrid is a new project involving the Universities of Frankfurt and Cambridge, Accelrys, IBM, and the Central Laboratory of the Research Councils in the UK (CCLRC), which is funded by the UK government, and aims to create a pilot database of materials properties (such as elastic stiffness, dielectric constants, optical properties, heat capacity, electronic band gap) based on quantum mechanical simulations.
Central to the project is the idea that the database should be dynamic. Consequently, the front-end, data and compute infrastructures will be linked so that calculations will be triggered automatically if gaps are found in the data. Rather than use proprietary file formats, we use CML for all internal communication, as this allows us to use any CML-aware code as the computational chemistry 'engine'.
In this contribution, I will discuss the MaterialsGrid project, paying particular attention to the role of CML. I will also discuss our initial ontological analysis, and the implementation details of our CML-aware version of CASTEP. Furthermore, work within our group on the development of CML dictionaries will be presented.

XML I/O Subsystem for Computational Nanoscience
Thomas Schulthess
Oak Ridge National Laboratory, United States
Coauthors: M. Summers, A. Kozhevnikov, F. Rose, J. Velev, A. Janoti
Abstract

Nanoscience is typical for science of the 21st century in that it is very interdisciplinary, combining traditional domains such as physics, chemistry, materials science, and biology. Computational nanoscience can build on highly sophisticated techniques that have been developed in the respective domains. The interdisciplinary nature of the field, however, requires scientists to apply very different computational methods and codes that have been developed independently over decades. Furthermore, computational nanoscience relies increasingly on leading high performance computing platforms that are becoming increasingly more complex, requiring specialization of software systems. Finally, software systems have to be highly configurable, since nanoscience is a rapidly developing field. In a nutshell, the following characteristics are typical for computational nanoscience:

- 1. Interoperability of codes that have been developed independently.
- 2. Use of customized highly optimized codes for particular computing architectures.
- 3. Flexibility and extensibility of the software system
- 4. Ability to incorporate legacy codes

From the point of view of data representations, this requires that input and output data formats are as unique as possible across different codes, or, at a minimum, defined in a way that they can be easily translated. The use of XML to express data and metadata is therefore a very natural and appropriate choice for many situations. However, in order to make XML based technology work in real application, we have to address several challenges:

- 1. Parsing and validation must be supported for various flavors of Fortran, in addition to C and C++, without the need for heroic rewriting of codes
- 2. The XML technology must be usable on high performance computing platforms.

- 3. Consensus on XML Schemas must be developed by code developers as well as users.

In this talk I will give an overview of the XML I/O subsystem, the tools, and the Schema development environment that we are implementing at ORNL. In the morning of Day 2 of the workshop, Anton Kozhevnikov will discuss how the XML I/O subsystem has been incorporated into the Stuttgart LMTO code and how this has enabled him to combined electronic structure calculations with model Hamiltonian simulations that use the ALPS toolkit (see also talk by Matthias Troyer in the same session). In a later talk on day 2, Michael Summers will give detailed introduction and demonstration of the XML I/O subsystem.
Research sponsored by the Laboratory Research and Development Program of ORNL, managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

Grid Enabled Molecular Science Through Online Networked Environments
Kim Baldridge
Uni Zurich, Organic Chemistry Institute, Switzerland
Coauthors: Karan Bhatia, Steve Mock, Jerry Greenberg
Abstract
GEMSTONE (Grid Enabled Molecular Science Through Online Networked Environments), is a cross-platform desktop application that provides users access to remote scientific web services, with emphasis in this work for computational chemistry and biochemistry applications. Gemstone is built upon the Mozilla foundation codebase and available as a Firefox extension, with automatic update of new releases. The new infrastructure is capable of providing a versatile desktop user interface to strongly typed web services for applications running in a distributed cluster computing environment. The user interface enables a user to save and reuse application parameters, with drag and drop capability on multiple platforms. The interface supports server-side dynamic user interface configuration update. First phase Gemstone implimentation provides support for applications such as Adaptive Poisson-Boltzmann Solver (APBS), Babel, GAMESS (General Atomic and Molecular Electronic Structure System), a visualization component, GARNET, as well as several small interface tools that have been rewrapped to be web services. Additional applications are in the works, being easily added using the Gemstone application support framework, an integrated framework for accessing grid resources that supports scientific exploration, workflow capture and replay, and a dynamic services oriented architecture. An outcome of this work has been the generation of an automatic web service wrapper, OPAL, which enables one to wrap legacy applications as Web services for deployment of applications as Web services in a matter of hours. This talk will show these efforts.

Automatically extracting metadata from CML for combinatorial simulations on the grid
Richard Bruin
Dept. of Earth Sciences, Univ. of Cambridge, United Kingdom
Abstract
As grid computing becomes more mainstream and scientists are able to submit larger number of calculations to be performed concurrently the difficulties faced by the scientist change dramatically. The difficulties move from ”where can I find a machine to run my job?” to being ” can run so many jobs now, how do I handle all of the data created?” I will describe how the eMinerals project uses CML program output for automatic collection of metadata. Our simulation codes (which include SIESTA and DL_POLY) generate three main types of CML output: metadata associated with the running of the program (eg program name, version number etc), parameters associated with the control of the simulation

(such as temperature or pressure, function cut offs etc), and key properties (such as average volume and final energy). We typically run these programs within a grid computing environment. I will describe the job submission procedures we have developed, which now include tools based on the Agent-X and RCommand libraries developed by CCLRC, and give some case studies.

On the structuring of the computational chemistry Virtual Organization COMPCHEM

Osvaldo Gervasi
Coauthors: Antonio Lagana, Antonio Riganelli
University of Perugia, Dept. of Maths and Comp Sci, Italy

Abstract
The first moves to structure the COMPCHEM (Computational Chemistry and Molecular sciences) virtual organization (VO) of the related sientific community are discussed. The efforts have been focused on the assemblage of a grid enabled molecular simulator (GEMS) by gathering together the human, hardware and software resources of several research laboratories. To this end the necessary suites of codes have been structured in a way to operate in a distributed coordinated way on an extended grid of computers. A particularly critical point has been found to be the sustainability of such an effort.

Specification of file formats for NANOQUANTA

Xavier Gonze
Universite Catholique de Louvain, Belgium
Coauthors: C.-O. Almbladh, A. Cucca, M. Marques, C. Freysoldt, V. Olevano, Y. Pouillon, M. Verstraete

Abstract
In order to allow softwares to interact and exchange data, file format specifications are mandatory. Widely agreed file format specifications are still lacking in the field of first-principles calculations of material properties.
One of the (numerous) objectives of the European Network of Excellence ”NANOQUANTA” is precisely to specify file formats, for the contents that are relevant to the scientific activity of its constituting nodes. After one year and a half of existence of NANOQUANTA, a large body of work has been done along these lines. It includes an inventary of the different contents, discussions of existing formats when relevant, as well as detailed (new) NetCDF specifications, for selected contents (density/wavefunctions). Pseudopotentials and geometry descriptions have also been examined. A document has been written, that gives specifications for the NANOQUANTA groups. It is hoped that the new specifications will be implemented in many different softwares, or (at least) will be the basis of even better file format specifications

A proposal for a unified norm-conserving pseudopotential format

Javier Junquera
Univ. Cantabria, Spain, Spain
Coauthors: Alberto Garcia and Matthieu Verstraete

Abstract
Pseudopotentials are widely used in electronic structure calculations and first-principles molecular dynamics simulations. At the present time, however, each computer code uses its own format for the input of pseudopotential information, which hinders interoperability and makes it difficult to compare

the results obtained with different programs. Here we present a proposal for a unified file format for norm-conserving pseudopotentials. The new format is based on XML, and thus platform independent and easily processed by a wide variety of off-the-shelf tools, including a new Fortran XML parser (see http://lcdx00.wm.lc.ehu.es/ag/xml). This means that first-principles codes using pseudopotentials can easily map the information in the XML file to their internal data structures. While the current proposal is limited to norm-conserving pseudopotentials, it offers a proof-of-concept for the general approach. Further discussion and development is ongoing under the auspices of the FSAtom project (http://dirac.cnrs-orleans.fr/fsatom_wiki/PseudoPotentials).

Data file exchange in Quantum-ESPRESSO

Paolo Giannozzi
Scuola Normale Superiore di Pisa , Italy

Abstract
The need to extract and process information (wavefunctions, charge density etc) from data files produced by electronic structure codes is strongly felt in the scientific community. In this talk I will describe the state of the art and the perspectives of data file exchange in the Quantum-ESPRESSO distribution. The goal is to make access to data files produced by QE codes as easy as possible. The ideal data file format should satisfy several different and sometimes conflicting requirements: - fast to read and write - easy to read, parse and write without special libraries - easy to understand, self-documented - portable across different computer architectures - portable across different software packages The solution we have devised uses a data directory, instead of a single file, thus exploiting the capabilities of the file system for data organization. Binary I/O is used for large records (wavefunctions, charge density) while smaller records (for instance, the structure and symmetry of the system, the pseudopotentials) are written to a XML file (formatted). A small specialized library (iotk, input/output toolkit written by G. Bussi) is used to write and read the XML files and records.

Data representation for biomolecular systems: technical and political challenges

Konrad Hinsen
Laboratoire Lé on Brillouin (CEA-CNRS), France

Abstract
Biomolecular simulations are characterised by complex data structures, large data sets, and the need to interface with equally complex experimental data. The existence of few big simulation packages with relatively closed user communities has so far discouraged any efforts towards interoperatibility, except for the representation of molecular structures, which are the main interface to experiment. I will give an overview of the data for which a standardized representation would be desirable, either for interoperability or for archiving, and discuss the criteria that should be taken into account when defining file formats. I will also report on recent attempts at reaching a concensus about interoperability and why they have failed so far.

References

[1] Hinsen, K. *The Molecular Modeling Toolkit: A New Approach to Molecular Simulations*, J. Comp. Chem. **21** 79-85 (2000)

Demonstration of XML tools: xmlf90

Jon Wakelin
University of Bristol, UK, United Kingdom
Coauthors: Toby White, Alberto Garcia
Abstract
XML is a common solution to problems of transferable data exchange, and code interoperability. However, current codes must be moved into the XML world. One option is pre/post-processing of existing formats, but a more direct method is integrating XML into the code itself. Many such scientific codes are written in Fortran - to this end xmlf90 is a modern Fortran library which facilitates XML input (parsing) and output. Parsing can be done within the SAX and DOM paradigms. In addition, xmlf90 has a specialized module for CML code to facilitate output of well-structured, semantically full CML documents, in a fashion non-intrusive and comprehensible to the scientific programmer unversed in XML technologies.

XSLT transforms
Toby White
Dept. of Earth Sciences, Univ. of Cambridge, United Kingdom
Coauthor: Jon Wakelin
Abstract
XSLT transforms allow the manipulation of XML data in a platform-independent, programming-language-agnostic manner. Here I illustrate two applications of XSLT transforms of real and practical use. Firstly, generation of useful, human readable, visual output from a non-application-specific CML file. A mixed-namespace XHTML/CML/SVG output file is shown, visualizable in any standards-conformant WWW browser, which includes 2D SVG graphs of quantities of interest, and 3D embedded visualizations (via Jmol) of molecular structures. (i) Of particular interest is the Jmol integration, which entailed additions to the Jmol codebase to allow Jmol to directly traverse the namespace XML document in which it is embedded and extract relevant CML data. (ii) Also of interest is the graph-drawing XSLT which, in a fully general fashion, will extract and draw 2D graphs of arbitrary scale.

XML Schema Design for First-Principles Molecular Dynamics
Francois Gygi
University of California, Davis, United States
Abstract

First-Principles Molecular Dynamics (FPMD) simulations are rapidly gaining importance in many areas of computational materials science, physics and chemistry. This is accompanied by a growing need to facilitate data exchange between FPMD simulation codes and post-processing tools or other simulation codes such as quantum Monte-Carlo codes. XML is emerging as the best candidate markup language for exchange of FPMD data. The use of XML must be supplemented with a definition of elements and attributes used to markup FPMD data. This can be done using the older Document Type Definition (DTD) syntax, or with the more powerful XML Schema language. We present an XML Schema definition of FPMD simulation data. This definition has been used successfully by the Qbox plane-wave, pseudopotential FPMD code for over two years. We discuss general rules that were followed in the design of the FPMD XML Schema specification. Semantic conflicts with other XML definitions are avoided through the use of XML namespaces. The use of a well defined XML syntax for FPMD simulation leads to improved reliability of simulation data since XML FPMD documents can be validated against their corresponding XML Schema definition and processed using extended stylesheet language transformations

(XSLT). These properties will be illustrated with examples taken from the use of the Qbox code.

Object representations for quantum simulations

Jeongnim Kim
NCSA, United States

Abstract
I present recent developments of object representations for quantum simulations in the context of Quantum Monte Carlo (QMC) simulations and tools to interface multiple applications. We have developed qmcPACK (Quantum Monte Carlo Package) in an object-oriented framework. qmcPACK exploits the natural mapping between the computational objects and the data representation in xml. I discuss QMC schema and QMC applications, focusing on i) the data representation to facilitate data exchange with other applications and ii) the execution model to perform QMC simulations with many run-time parameters.

The ALPS project

Matthias Troyer
Theoretische Physik, ETH Zü rich , Switzerland

Abstract
The ALPS project (Algorithms and Libraries for Physics Simulations) is an open source effort aiming at providing high-end simulation codes for strongly correlated quantum mechanical systems as well as C++ libraries for simplifying the development of such code. ALPS strives to increase software reuse in the physics community.

Unified XML I/O approach in DFT and model codes for real materials: from crystal structure to magnetic susceptibility

Anton Kozhevnikov
Russian Acad. Sciences, Russia, Russian Federation

Abstract
We report progress in implementing XML I/O in the TB-LMTO-ASA band structure code. Parsing and validation of xml source as well as modification and saving of DOM tree were done with the Xerces/Xalan libraries. The XML I/O paradigm was then used in calculations of Heisenberg exchange parameters. These parameters were passed to ALPS thus completing the chain ”crystal structure -⟩ one-electron Hamiltonian -⟩ model parameters -⟩ megnetic susceptibility”.

Data representation in the Molpro quantum chemistry package

Peter Knowles
Cardiff University, UK, United Kingdom

Abstract
The Molpro package (http://www.molpro.net), which is focused on ab initio electronic structure computations for small and medium-sized molecules, contains several facilities for the import and export of portable structured data. Export of molecular geometries and wavefunction data follows an extended

CML format, and is augmented by a fully-tagged transcript of an entire job. Internally, the program makes use of a xml-expressed library of basis sets and pseudopotentials that is accessible through interactive documentation, and that is capable of transformation to other formats. These features will be demonstrated through a number of examples, which will then form the basis for a discussion of outstanding and difficult issues, and of the potential for interfacing to other codes and representation schemes.

A common format for Quantum Chemistry

Elda Rossi
CINECA, Italy

Abstract

A Common Format for Quantum Chemistry has been designed and implemented. It responds to the requirement of communication between different QC programs and aims at enhancing code interoperability. A general discussion about a possible model for data organization in the QC domain is reported: small data describing geometry, basis set and symmetry are organized within an XML based format (QC-ML), while large binary data describing integrals and coefficients are organized within an HDF5 based format (Q5COST). Some applications that use the proposed format for running a complex chain of heterogeneous programs, both general purpose and home made, will be presented as well. Last, we will discuss a possible follow up of this activity in terms of workflow and distributed computations (Grids and Web Services).
This activity has been carried out within the COST in Chemistry D23 project ”MetaChema”, in the Working Group ”A meta-laboratory for code integration in ab-initio methods”.

Building an Infrastructure for Quantum Chemistry: Data Sharing and Graphical User Interfaces

Sherwood Paul
Daresbury Laboratory, UK, United Kingdom
Coauthor: Phil Couch

Abstract

We will discuss the work that has been carried out under the CCP1 (Collaborative Computational Project No. 1) with the objective improving interoperability between quantum chemical codes, and providing common tools such as graphical interfaces. I will discuss the issues that have arisen in the consultation processes and discuss possible ways of progressing further the definition of standards for data exchange. I will describe an open source project which aims to provide a common user environment to a number of QM and other modelling codes, and which is being used as prototype in exploring some of the data exchange issues.

Material leading to discussion: Data markup hierarchies and interoperation

Peter Murray-Rust
Unilever Ctr for Mol. Informatics, U. of Cambridge, United Kingdom

Abstract

I would like to give an overview of markup and to review current approaches. My current interpretation is very roughly that there is a rough hierarchy - content+syntax, dictionaries, semantics.
CML tackles content+syntax and provides XSD-like validation (there are a few validation extensions).

CML aims to:

- provide unique labels for agreed concepts usable in a wide range of domains
- provide validation
- provide common programmatics functionality
- act as a program- and laboratory- independent nucleus on which others can build.

There seem to be the following types of markup:

- content extensions because the concept is not and cannot be realised in CML. An obvious example is pseudopotential (though not basisSet, which CML supports).
- Program-specific markup (e.g. GAMESS). In CML-language this is a mixture of content and dictionary. It looks as if it can be automatically transformed to CML and a dictionary.
- semantic languages, such as Agent-X, dREL (IUCr), OMDOC, and unstructured RDF. CML itself is neutral to these and can be used by any.

An Introduction to the Use of XMLDIR and QuickSchema
Michael Summers
Oak Ridge National Laboratory, USA, United States
Abstract
XMLDIR has been developed as a common IO subsystem for a family of scientific codes. Its development started with an initial system and continues through incremental refinements based on experience gained by incorporating XMLDIR within existing codes. This experience has provided us with insights into the techincal and organization requirements of such a common IO susbsystem. These insights include an understanding of:

- 1. what new users need to know and do as they integrate XMLDIR into their codes, and
- 2. the need for an easy to use, web-based, collaborative, schema development tool.

This talk will provide a quickstart tutorial for the XMLDIR system and an associated demonstration of a schema development tool.
Research sponsored by the Laboratory Research and Development Program of ORNL, managed by UT-Battelle, LLC for the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

XML and the Vienna Ab-initio Simulation Package (VASP)
Orest Dubay
Univ. of Vienna, Austria
Abstract
Our experience shows that a XML based format is a very successful concept in VASP, though it has also some limitations. VASP uses a XML-based output format for more than two years. As a prove of the concept, XML was also used for creation of batch-jobs starting multiple VASP calculations. Our

initial goals, design decisions and achievements will be discussed to show the strengths as well as some shortcomings related to the use of XML in VASP.
The VASP XML output can be processed by a tool called ”p4vasp” which will be presented both from the user and the developer perspective. P4vasp can create plots of various quantities (e.g. the density of states and the band-structure), show iso-surfaces, manipulate structures and create simulated scanning tunneling microscopy images. P4vasp is very modular, it is implemented in c++ and python and it can be used as a library in simple python scripts.
At the end we will introduce our vision of the sharing and propose a way how to assemble a data-sharing system from the available open-source technologies.

Information Exchange in Computational Chemistry and Materials Physics - AgentX

Phil Couch
CCLRC-Daresbury Laboratory, United Kingdom
Coauthors: Rob Allan, Paul Sherwood, Peter Knowles

Abstract
Increasingly, solving complex scientific problems requires the use of more than one application cooperatively. Considerable effort has been invested in developing general tools that can be used to specify and execute computational workflows that address such problems. However, despite recent advances in this area, it remains difficult to exchange information between applications, principally due to a lack of data standards.
The problem of information exchange derives, in part, from a difficulty in finding the context or meaning of data represented in documents. This can be addressed through the development of common representations for computational data. A common data model specifies how data should be structured and described according to its context. Tools developed to work with data that conform to this model can form a common component of applications exchanging information. The problem can also be addressed by developing standards that can be used to specify how to locate data with a particular context in a document. The AgentX framework is being developed to manage this task. AgentX abstracts away from the format of the data, resulting in the ability to work transparently with data conforming to different standards. The user is not required to understand the underlying data format. AgentX takes the form of a library, presenting a simple interface and being accessible via wrappers to applications written in C, Python, Perl and Fortran.
The general principles of the AgentX framework will be discussed, along with some current applications in the computational domain.

Evolutionary Construction of Higher-level Constraints over an XML Language; Formalizing Informal Semantics

Toby White
Dept. of Earth Sciences, Univ. of Cambridge, United Kingdom
Coauthor: R. Bruin

Abstract
Within the eMinerals project, we have access to a number of computational chemistry codes; at different levels of theory, and with different purposes; which have in common the notion of simulating the interactions of a system of atoms. We have worked to give these codes CML output, with the initial aim of facilitating the tasks of data interchange. In addition, we have developed visualization tools, based on XSLT, which present the output of the codes in a fashion useful to the simulation scientist.

These tools are designed to work agnostically of the code of origin; so the same visual output can be seen from any code which outputs correctly organized CML. The development and refinement of these visualization tools has, however, highlighted the existence of informal semantics above and beyond the mere syntactical requirements of CML, and has forced us to address these in increasingly formal ways. The specification of these semantics, which effectively subsets and constrains CML further, is of wider applicability than mere visualization.
This talk will discuss the use of XSLT as a transform language for CML, and demonstrate the use of the visualization tools. From this, a wider theme will emerge about the necessity and utility of constraints beyond the syntactical validitiy of CML.

CCPN - automatic code generation from UML data models

Rasmus Fogh
University of Cambridge, United Kingdom
Coauthors: Boucher, W., Vranken, W., Stevens, T.J., Pajon, A., Ionides, J., Henrick, K., Laue, E.D.

Abstract
Software interoperability, data exchange and data harvesting require a data standard that is comprehensive, general, and covers intermediate as well as final results. For practical use it must be precisely defined, easy to validate against, and require few resources to maintain. To become widely adopted it must be attractive to programmers with widely different tastes.
CCPN presents the Memops data modeling framework, developed as part of our work on data standards and software integration for macromolecular NMR spectroscopy, as our solution to this problem. Memops begins with an abstract, implementation-independent data model to be written in UML (unified modeling language). Automatic code generation then produces fully functioning data access implementations for a variety of languages (Python, Java, C/C++, Perl) and storage implementations (XML file and relational database), directly from the abstract model. The autogenerated libraries include object-oriented API implementations, I/O mappings and I/O code (for files) or persistence layers (for databases), XML and database schemas, and documentation. All code works seamlessly together, so that e.g. data loading is triggered automatically when the data are needed. Work on client/server support, transactions/rollback and access control is in progress. The API code verifies the validity of all input data according to the constraints of the model; arbitrarily complex constraints can be entered during of the modeling process. A notifier/listener system is in place to trigger external events when the data change. The model is organised in separate packages, making it possible to cover a wide area of data while allowing each application to use only the relevant parts of the model.

Handling of large datasets in Quantum Chemistry

Antonio Monari
Universita di Bologna Italy
Coauthor: S. Evangelisti

Abstract
Our group has been involved in the development of a common format for code interoperability and communication in the field of Quatum Chemistry. Two libraries have been developed, for the treatment of ”small” and ”large” data sets, respectively. In particular, we have focused our work on the problem of ”large” (binary) data, that represent a peculiarity of QC. We have implemented developed a FORTRAN library based on the HDF5 format (Q5COST), that permits data transfer between the programs that produce one and two-electron integrals (at the moment, DALTON and MOLCAS), and chains for

the post-SCF treatment (CI or PT algorithms developed in Toulouse and Ferrara). The use of HDF5 has shown some important advantages: high efficiency for I/O operations on large binary files, independence from the architecture of the resulting files. At the same time, the choice of FORTRAN makes the library of easy use within the community of QC program developers.

Input handling in CP2K

Fawzi Roberto Mohamed
Scuola Normale Superiore, Pisa, Italy
Coauthors: Teodoro Laino, Joost VadeVondele

Abstract
The f90 opensource CP2K project http://cp2k.berlios.de has just rewritten the input. The new input structure has been inspired by XML, and is described in a declarative way. This enables the automatic generation of an always up-to-date description of the input, and the automatic handling of the input. Along with the documentation, the usefulness of our approach has been demonstrated by the ability automatically generate inputs (used in montecarlo, path integral,...).

Participant List

Osvaldo Gervasi (osvaldo@unipg.it)
University of Perugia, Dept. of Maths and Comp Sci Italy

Rasmus Fogh (r.h.fogh@bioc.cam.ac.uk)
University of Cambridge United Kingdom

Antonio Monari (amonari@fci.unibo.it)
Universita' di Bologna Italy

Thomas Schulthess (schulthesstc@ornl.gov)
Oak Ridge National Laboratory United States

Alberto Garcia (wdpgaara@lg.ehu.es)
Universidad del Pais Vasco, Bilbao Spain

Philip Couch (p.a.couch@dl.ac.uk)
CCLRC United Kingdom

Walter Temmerman (W.M.Temmerman@dl.ac.uk)
Daresbury United Kingdom

Martin Lueders (m.lueders@dl.ac.uk)
Daresbury Laboratory United Kingdom

Fawzi Roberto Mohamed (fawzi@gmx.ch)
Scuola Normale Superiore, Pisa Italy

Markus Holzmann (markus@lptl.jussieu.fr)
LPTL, Jussieu, Paris France

Jon Wakelin (Jon.Wakelin@bristol.ac.uk)
University of Bristol, UK United Kingdom

Xavier Gonze (gonze@pcpm.ucl.ac.be)
Universite Catholique de Louvain Belgium

Sherwood Paul (p.sherwood@dl.ac.uk)
Daresbury Laboratory, UK United Kingdom

Paolo Giannozzi (p.giannozzi@sns.it)
Scuola Normale Superiore di Pisa Italy

Javier Junquera (javier.junquera@unican.es)
Univ. Cantabria, Spain Spain

Peter Knowles (KnowlesPJ@Cardiff.ac.uk)
Cardiff University, UK United Kingdom

Peter Murray-Rust (pm286@cam.ac.uk)
Unilever Ctr for Mol. Informatics, U. of Cambridge United Kingdom

Francois Gygi (fgygi@ucdavis.edu)
University of California, Davis United States

Kim Baldridge (kimb@oci.unizh.ch)
Uni Zurich, Organic Chemistry Institute Switzerland

Toby White (tow21@cam.ac.uk)
Dept. of Earth Sciences, Univ. of Cambridge United Kingdom

Richard Bruin (rbru03@esc.cam.ac.uk)
Dept. of Earth Sciences, Univ. of Cambridge United Kingdom

Matthias Troyer (troyer@itp.phys.ethz.ch)
Theoretische Physik, ETH Zü rich Switzerland

Konrad Hinsen (khinsen@cea.fr)
Laboratoire Lé on Brillouin (CEA-CNRS) France

Stefano Evangelisti (stefano@irsamc.ups-tlse.fr)
Universite Paul Sabatier France

Elda Rossi (e.rossi@cineca.it)
CINECA Italy

Jeongnim Kim (jnkim@ncsa.uiuc.edu)
NCSA United States

Orest Dubay (orest.dubay@univie.ac.at)
Univ. of Vienna Austria

Daniel Wilson (wilson@kristall.uni-frankfurt.de)
Univ. of Frankfurt Germany

Michael Summers (summersms@ornl.gov)
Oak Ridge National Laboratory, USA United States

Anton Kozhevnikov (anton@imp.uran.ru)
Russian Acad. Sciences, Russia Russian Federation