Proceedings of the UK e-Science All Hands Meeting 2006, © NeSC 2006, ISBN 0-9553988-0-0

Providing an Effective Data Infrastructure for the Simulation of Complex Materials

L. Roberts, L.J. Blanshard, K. Kleese van Dam

CCLRC eScience Centre, Daresbury Laboratory, Warrington, WA4 4AD

S. L. Price, L.S. Price and I. Brown

Department of Chemistry, University College London, 20 Gordon Street, London, WC1H 0AJ

Abstract

CCLRC have developed a suite of data management tools for the Engineering and Physical Sciences Research Council (EPSRC) funded e-Science project 'The Simulation of Complex Materials' [1], which ran from 2002 - 2005. The focus of the project was to aid the development of a computational technology for the prediction of the polymorphs of an organic molecule prior to its synthesis, which would then provide the ability "*to control the unwanted appearance of polymorphism and to exploit its benefits in the development, manufacture and processing of new molecular materials"* [2]. Prior to the project the data of interest was distributed across a multitude of sites and systems, with no simple formal methods for the management or distribution of data. This was considerably hindering the analysis of the results and the refinement of the prediction process and it was therefore essential to rationalise the data management process. The initial concern was for the collection and safe storage of the raw data files produced by the simulations during the computation workflow [3]. This data is now stored in a distributed file system, with tools provided for its access and sharing. As the data was not annotated with metadata it was difficult for it to be discovered and reused by others and so web interfaces were implemented to enable the cataloguing of data items and their subsequent browsing. In addition there was no fine grained access to the data. Specific crystal data is now parsed from the simulation outputs and stored in a relational database, and a web application has been deployed to enable extensive interrogation of the database content. This paper will elaborate on these tools and describe their impact on the achievement of the project's aims.

Background

A crystal may have different polymorphs, (different arrangements of the molecules in the crystal lattice), and

"different polymorphs have different physical properties, and so there are major problems in quality control in the manufacture of any polymorphic organic material. For example, a polymorphic transformation changes the melting point of cocoa butter and hence the taste, the detonation sensitivity of explosives producing industrial accidents, and the solubility changing the effective dose of pharmaceuticals." [2].

Therefore a method of predicting which crystal structure a given organic molecule will adopt under different conditions would have considerable benefit in product development across the range of molecular materials industries.

The computational chemistry group at UCL have developed computational methodologies for predicting the energetically feasible crystal structures of small, rigid, organic molecules [4]. Each simulation involves the running of multiple programs to generate these crystal structures, at considerable computational and human expense.

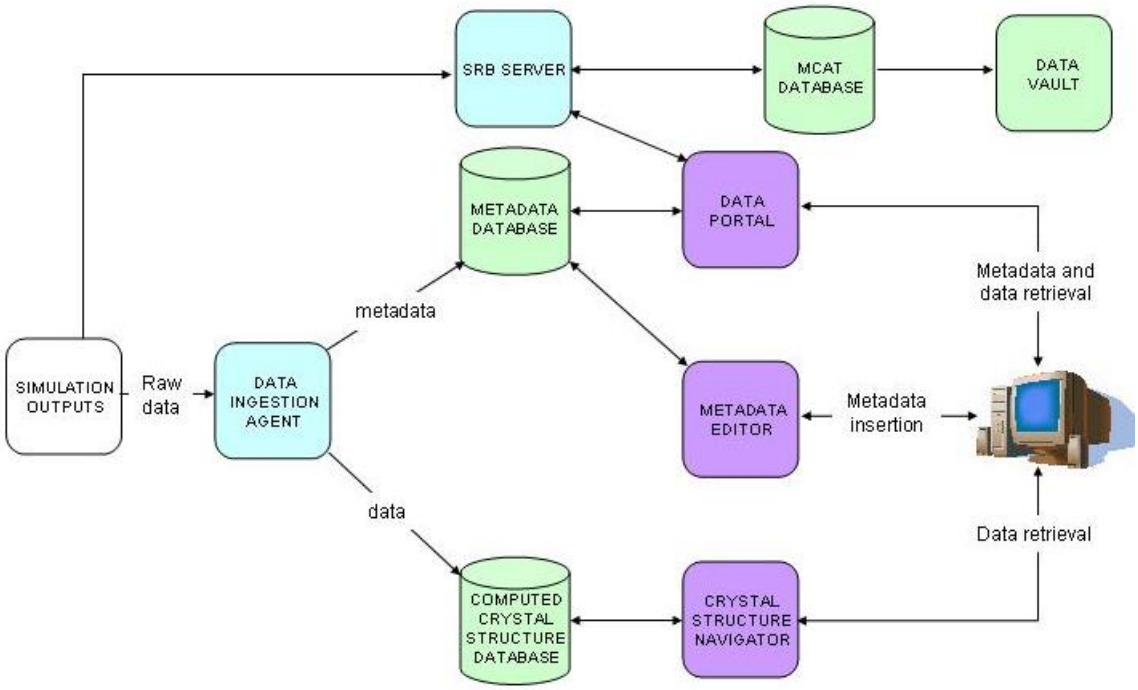

Figure 1: Deployed architecture for the eMaterials project

The process produces a great quantity of heterogeneous data estimating various mechanical, morphological and spectroscopic properties of the computed crystal structures. However there was little support for accessing and managing this data and this hindered the project's progress and collaborative efforts. To counter this CCLRC developed an effective data management infrastructure to facilitate the creation, storage and analysis of data and metadata.

Architecture

This infrastructure, as shown in figure 1, enables a more efficient cycle of discovery, analysis, creation and storage of both data and metadata. The tools provided for the management and discovery of metadata are the Data Portal [5], the Metadata Editor [6], the CCLRC metadata schema [7] and the metadata database [8]. The tools for the management of raw and processed data are the Computed Crystal Structures Database, the Crystal parser, the Crystal Structure Navigator and the Storage Resource Broker (SRB) [9]. The SRB was developed by the San Diego Super Computing Centre (SDSC).

After a simulation run the results are uploaded to the relevant storage media and can be discovered and shared with other scientists - for example, to use for comparison when a new polymorph is discovered experimentally that has already been computationally predicted. [10]

From the simulation outputs the raw data is processed in three ways:

i. A subset of the data is parsed from the files and inserted into the crystal database.
ii. The data files are uploaded into SRB.
iii. Metadata about the data is entered into the metadata database using the metadata editor.

Data Portal can then be used to discover and retrieve raw data from the SRB, and the Crystal Structure Navigator can be used to extract and compare data on the crystal structures from the computed crystal structure database.

Storage Resource Broker (SRB)

The first issue to be overcome was the lack of tools to manage the collection and safe storage of simulation outputs, which also made it difficult to share data. Data files from simulation runs were generally located on users' machines or on the machine on which the computation had been run.

In 2002 SRB was chosen as the middleware because of its appropriateness and CCLRC's strong links with SDSC. GUIs (Data Portal and Metadata Manager) that used the Scommands behind the scenes were written to allow users access to the grid enabled storage resources.

SRB is a distributed file system that allows files to be shared across multiple resources whilst providing access through a single interface and showing a unified logical view of the virtual organisation's data. As such data is organised in a logical rather than a physical context, and the complexity of locating and mediating access to individual data items is abstracted away from the users. Access control is simply implemented with the data originator able to control and configure access permissions to their data for other project members. The raw data produced by the simulations is now stored on a number of distributed resources managed by the SRB.

Metadata Database

Although the use of SRB facilitates data distribution and replication, the value of the data is in its discoverability – and the annotation of the data holdings with metadata is what achieves this.

The raw data files in SRB are now catalogued with metadata, and this data is stored in the metadata database which uses the CCLRC metadata schema. This is a common model that has been re-used in other CCLRC projects that required metadata holdings. As the same logical model has been re-used across different projects the applications written against this model, (Data Portal and Metadata Manager), have also been reused on other projects, with very little customisation required.

Raw data files are grouped into datasets, and datasets are grouped under studies with a name, start date, end date and originator details. The studies are also linked to topics, keywords and investigator details. The database tables for the data file and data set entities also store the physical location of the data in SRB, alongside the metadata.

Metadata Manager

The CCLRC metadata manager is a web-based tool for the insertion and manipulation of entries in the metadata database. Typically users annotate their datasets and data files with details such as the provenance of the data and its location in SRB. Users can organise their data by creating and editing information about studies and then adding datasets and data files to the hierarchy. The metadata forms the basis for the search and retrieval of data by the Data Portal.

In the most recent version of Metadata Manager users can also add topics, so removing the last vestiges of a centrally controlled vocabulary in the infrastructure. Scriptable command line tools for the insertion of metadata have also now been written.

Data Portal

The CCLRC Data Portal is a web-based front-end that enables the scientists to browse the metadata database and discover data resources from physically diverse institutions. Data is displayed in a virtual logical file system – so files are displayed by topic (such as the molecule) rather than by geographical location, this being of much more use to the project participants. The required data resources can then be downloaded from SRB directly to the users' machines.

The Computed Crystal Structure Database

Although SRB allowed the searching of related data it did not provide any tooling for data comparisons and refinement as there was no fine grained access to the data, which was still stored in unstructured text files. This hindered analysis of the results and it was very difficult to look at, for example, all the crystal structures with a total lattice energy beneath a certain threshold. Therefore the project requested a bespoke database (figure 2) specific to their requirements

and data, which would enable them to drill down and query across data.

The CCLRC Computed Crystal Structure Database is a relational database running on Oracle 10G and is hosted on the National Grid Service (NGS). It is used to store specific data about crystal structures, rather than metadata or unprocessed raw data in a file or CLOB format. The data model for the calculated crystal database was designed to allow easy interrogation of the database and provide good performance for complex queries as the database grew.

There are five tables for representing the crystal data –

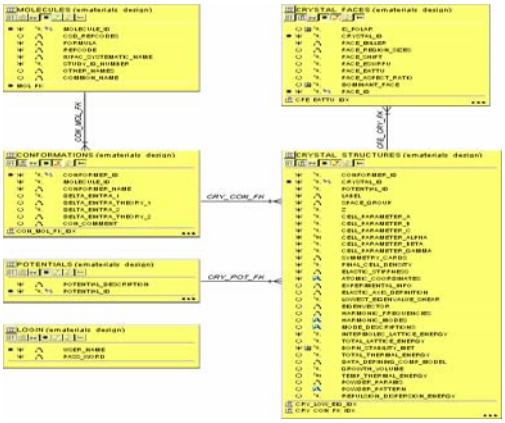

Figure 2: Crystal database schema

molecules, conformations, potentials, crystal_structures, and crystal_faces. At the moment it is quite a small database – holding just 42 molecules and around 2493 crystal structures with 228788 crystal faces between them. The attributes stored are molecule, conformation, potential and crystal descriptors, as well as growth volume attributes, second derivative properties and various energies. However, the database is being continually expanded as more molecules are studied, and the range of properties to be stored can be extended in the future if necessary. As more data has been added the database has scaled well and the applications have not suffered a drop in performance.

Populating the Database

An automated data ingestion agent, (written by Ian Brown at UCL), parses individual datum from the fixed format output files from the simulation runs and populates the computed crystal structure database with a subset of the simulation data. Some extra information about the molecule or conformation or potential has to be entered into a configuration file by the project administrator (Louise Price) and the script is then run from the command line. If any information on the study being uploaded is already contained in the database it is updated otherwise new records are inserted into the database.

Crystal Structure Navigator

The crystal structure navigator is a powerful query tool that provides users with the ability to search for crystal data which fits certain criteria by allowing the construction and execution of queries against the calculated crystal database, (figure 3). The application runs on a Tomcat web server and is written with JSPs and servlets. It has a simple and intuitive interface from which the user can compile their display and search options.

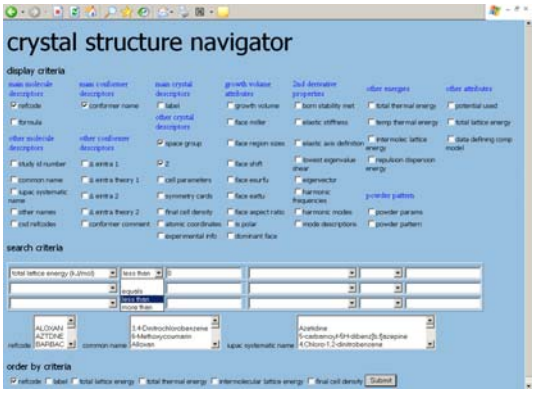

Figure 3: Crystal Navigator selection screen

From the display criteria section of the web page users can choose which combination of any of the forty-eight properties held in the database they would like to display. The search criteria section allows up to six criteria to be composed – with the user choosing the property, the relational operator and the value - such as 'refcode like E-N-I' or 'total lattice energy <-100'. To prevent user errors and simplify selection the relational operators are only enabled after the user has selected a property – for example if a user selects a textual property, such as common name, it would not make much sense to offer them 'more than' or 'less than' as operators; and likewise if they picked a numerical field such as total_lattice_energy then the

comparative operator 'like' would not be offered to them. There is also a section of 'quickpicks' – which shows a non-static list of values for the attributes refcode, common name and iupac_systematic_name for the user to choose from. This list is updated automatically as new entries are made in the database. Lastly the user can choose to impose an ordering on the results – for example the results can be ordered alphabetically by refcode, or by total_lattice_energy ascending.

Thus the complexity of relating the data fields, joining the relational tables and restricting the set of data to be returned is all hidden from the user whilst most of the functionality of SQL is made available in a user friendly interface. The results are then displayed on a results page, with only those data that fit the user's criteria and only those properties that the user selected being displayed. The user can download the results into a spreadsheet and this is now being used for the primary scientific analysis and for publications, and can also be used as the basis for input into other simulations. If the user wishes to perform a new search they can bring up a new criteria selection page, or they can review the criteria they just submitted.

Conclusion

The project has supplied an effective infrastructure for the management and utilisation of data. The provided suite of software is used daily to aid the computational studies on the polymorphism of various molecules being performed by the “Control and Prediction of the Organic Solid State” [2] project of the Research Council UK’s Basic Technology Program. The project uses the tools to make comparisons across the range of molecules being studied to refine and develop techniques for polymorph prediction. The feedback received from the scientists on the project has been extremely positive and the software suite is now essential for the storage, discovery and analysis of their data. The components of the architecture that deal with metadata have been reused on other projects with only a small and easy amount of customisation has been required.

References

[1] Simulation of Complex Materials project http://www.e-science.clrc.ac.uk/web/projects/complexmaterials

[2] Brief Overview, Control and Prediction of the Organic Solid State http://www.cposs.org.uk

[3] Blanshard, L.; Tyer, R.; Kleese van Dam, K. “eMaterials: Integrating Grid Computation and Data Management Services”; UK e-Science All Hands Meeting, 2004, Nottingham, UK.

[4] Price, S. L. "The Computational Prediction of Pharmaceutical Crystal Structures and Polymorphism." *Adv. Drug Deliver. Rev* **2004**, *56*, 301-319.

[5] Drinkwater, G.; Kleese van Dam, K.; Manandhar, A.; Sufi, S.; Blanshard, L. “Data Management with the CCLRC Data Portal”; International Conference on Parallel and Distributed Processing Techniques and Applications, 2004, USA.

[6] CCLRC Metadata Editor http://www.e-science.clrc.ac.uk/web/projects/scientific_metadatamgnt

[7] CCLRC Scientific Metadata schema http://www.escience.cclrc.ac.uk/documents/staff/shoaib_sufi/csmdm.version-2.doc

[8] Blanshard, L.; Kleese van Dam, K.; Catlow, C. R. A.; Price, S. L. “Simulation of Complex Materials: Database Design for Metadata”; UK e-Science All Hands Meeting, 2003, Nottingham, UK.

[9] SRB Home Page http://www.sdsc.edu/srb/index.php

[10] Vishweshwar, P.; McMahon, J. A.; Oliveira, M.; Peterson, M. L. & Zaworotko, M. J. "The Predictably Elusive Form II of Aspirin." *J. Am. Chem. Soc.* **2005,** *127*, 16802-16803