ENABLING EFFECTIVE COLLABORATION THROUGH A WEB-ENABLED DATA INFRASTRUCTURE

L. Roberts, L.J. Blanshard, R.P Tyer, K. Kleese van Dam
CCLRC eScience Centre, Daresbury Laboratory, Warrington, WA4 4AD, UK
l.e.c.roberts@dl.ac.uk, l.j.blanshard@dl.ac.uk, r.p.tyer@dl.ac.uk, k.kleese-van-dam@dl.ac.uk

ABSTRACT

The CCLRC is one of Europe's largest multidisciplinary research organisations supporting scientists and engineers world-wide. It operates world-class large-scale research facilities, provides strategic advice to the government on their development and manages international research projects in support of a broad cross-section of the UK research community. CCLRC is involved in the development of grid and data management tools for a number of e-Science projects in the UK including the Engineering and Physical Sciences Research Council (EPSRC) funded e-Science project 'The Simulation of Complex Materials' or *e-materials* [1]. The focus of the multi-institution project was to aid the development of a computational technology for the prediction of the polymorphs of an organic molecule prior to its synthesis. The project's use of grid computation and workflows has enabled the running of very large scale computations and simulations, with the subsequent production of large amounts of scientific data. In order for the simulation data to be used and shared effectively the project required an effective data infrastructure to ease data discovery for both data originators and their collaborators. As such the data needed to be self-reflexive and organised into an intuitive and flexible hierarchy; and access to the data to be possible from a variety of perspectives. The annotation of the data with metadata, the usage of relational databases and the deployment of web-enabled applications has enabled the browsing and discovery of data and metadata and underpinned the collaborative nature of the project.

KEY WORDS

e-materials, modelling, database, data sharing, polymorph prediction

1. Introduction

The *e-materials* project brought together a number of chemists from the Royal Institution, London and University College London and computer scientists from UCL and CCLRC. Initially they were running computational codes such as DMAREL [11] sequentially on any machine they had available. However their simulations across a set of parameters and their complicated workflow meant that it would take several months to complete the study of a particular molecule. In addition, their data consisted of the textual output from their many simulation runs. As a result it was extremely time-consuming to collate the results into a spreadsheet so that they could produce graphs and study any relationships in the data. Furthermore, the data was distributed across a multitude of individuals, institutions, sites and systems, with no methodology for the collection, storage or retrieval of data. This was considerably hindering the collaboration's analysis of their data and hence making it extremely difficult to further refine the prediction process. As such it was essential to rationalise the data management process in a way that would make the sharing of data both easy and implicit.

UCL Computer Scientists implemented a grid solution to run the simulations in parallel on a cluster of machines [12]. Also RI chemists made use of a Condor pool at UCL to run large numbers of simulations.

In terms of data management and sharing, the first priority was to create a mechanism for the collection and safe storage of the raw data files produced by the simulations during the computation workflow [2]. This raw data is now collated and stored in a distributed file system, which has tooling provided for the secure access and sharing of data between project members.

However, due to the sheer quantity of raw data and the non-description of the data in any way, it was difficult for it to be discovered and reused by other project members or even the data originator. Therefore a web interface was implemented to enable the cataloguing of data items with metadata, and another application was written to enable browsing of data via their metadata attributes.

Finally, there was no fine grained access to the data and it was very difficult to compare across data sets. Specific crystal data is now automatically parsed from the simulation output files and written to a relational database, and a web application was 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.

1.1. Scientific Drivers

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." [3].

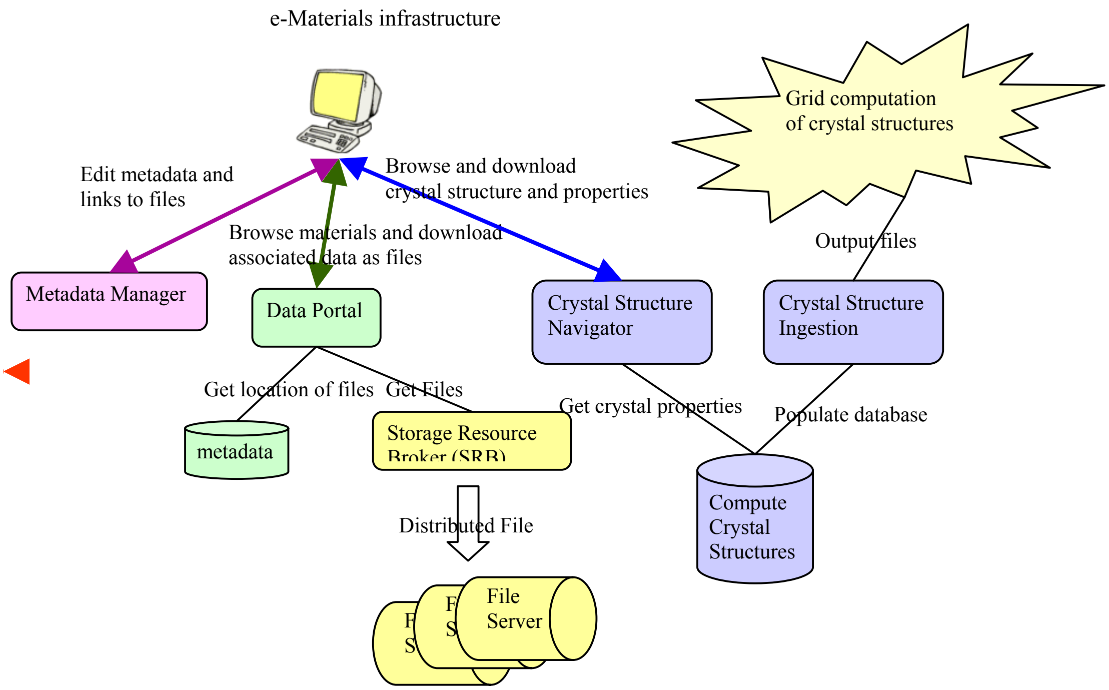


Figure 1: Architecture deployed for the project.

Therefore a method of polymorph prediction would be extremely valuable in aiding product development in many 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 prediction involves the running of multiple simulations to generate these crystal structures.

This process produces a great quantity of heterogeneous data estimating various mechanical, morphological and spectroscopic properties of the computed crystal structures. However, prior to the project 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 and a suite of web-enabled applications to facilitate the creation, storage, retrieval and analysis of data and metadata.

2. Infrastructure

This infrastructure, (as shown in figure 1), enables a much more efficient cycle of discovery, analysis, creation and storage of both data and metadata for project members across diverse institutions.

The tools provided for the collection and storage of data and metadata are the CCLRC metadata schema [5], the metadata database [6], the Metadata Manager [7], the Computed Crystal Structures Database, and the Storage Resource Broker (SRB), which was developed by the San Diego Super Computing Centre [8]. Data can be discovered using the Data Portal [9] and the Calculated Crystal Navigator.

When a simulation has finished running the results are uploaded to the relevant storage media where they 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]

So, from the simulation outputs the raw data is processed in three ways:

- the output files are parsed and a subset of the data is written to the calculated crystal database. The simulation codes that produce the output are used generally across the scientific community and the scientists are only interested in particular values.
- the output files themselves are written to SRB
- Metadata about the files, including their location in SRB, are entered manually into the metadata database using the Metadata Manager.

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.

2.1. Storage Resource Broker (SRB)

The first issue to be addressed 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 user's machines or on the machine on which the computation had been run. The raw data produced by the simulations is now stored on a number of distributed resources which are managed by the SRB.

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; this meaning that user's can retain control over who has access to their data.

2.2. 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. 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.

A study represents the top level of work undertaken by the individual or project – such as 'paracetamol'. Under this study will sit various datasets on an aspect of the study. For example a dataset may be comprised of a set of the crystal structures of the lowest energy structures found on the simulation run, or a collection of input files used for the minimisation and second derivative property calculations for the structures in the first dataset. The data object table entity is then used to represent the lowest level of data storage and is a single file or logical grouping of files that are physically stored and retrieved as one logical entity.

Whilst the data objects exist in the real world as artifacts in SRB, the concepts of study and dataset are abstract notions created by the experimenter and as such offer them the flexibility to impose a hierarchy on their work that meets their specific requirements to reflect the breakdown of their work – e.g. users can just interact with the metadata model at the study level or at all three levels.

2.3. Metadata Manager

The CCLRC Metadata Manager (figure 2) 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 which has been attached to the data then forms the basis for the search and retrieval of data by the Data Portal.

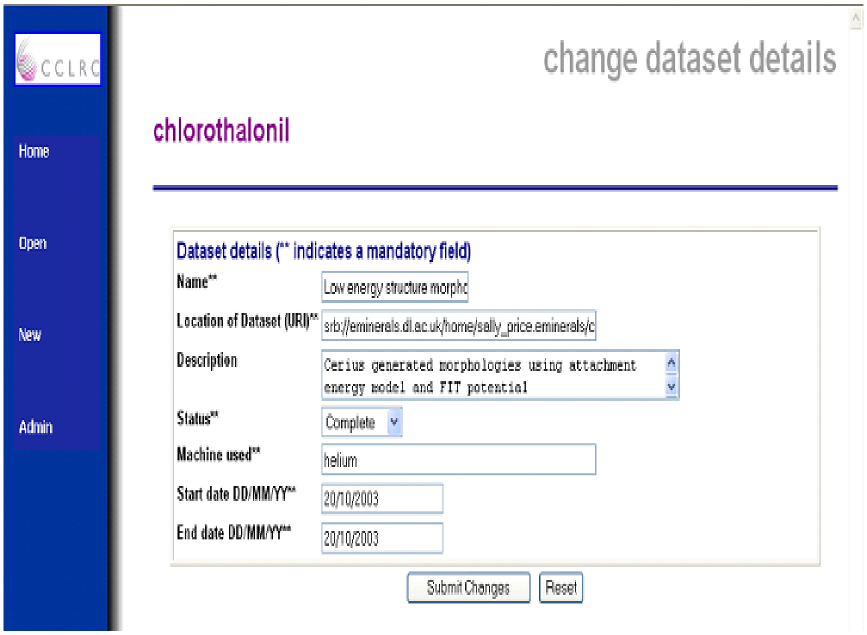
2.4. Data Portal

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 information is displayed by logical grouping such as study or topic (such as the molecule) rather than by geographical location, this being of much more use to the project participants. The users are initially shown a study view, and from this they can drill down to their desired level of granularity, and the required data resources can then be downloaded from SRB directly to the users' machines if desired.

2.5. Computed Crystal Structure Database

Although SRB allowed the searching of related data it did not provide any tooling for data comparisons or refinement as there was no fine grained access to the data - which was still stored in unstructured text files in SRB. This hindered collaborative efforts and the analysis of the results and it was very difficult to extract, for example, all the crystal structures with a lattice energy beneath a certain threshold.

The 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. The data model for the calculated crystal database was designed to allow easy interrogation of the database and provide good performance for complex queries. There are five tables for representing the crystal data: molecules, conformations, potentials, crystal_structures, and crystal_faces. At the moment it is quite a small database – holding just 34 molecules and around 1900 crystal structures; and 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.



change dataset details

chlorothalonil

Dataset details (** indicates a mandatory field)

Name**	Low energy structure morpho
Location of Dataset (URI)**	srb://eminerals.dl.ac.uk/home/sally_price.eminerals/c
Description	Cerius generated morphologies using attachment energy model and FIT potential
Status**	Complete
Machine used**	helium
Start date DD/MM/YY**	20/10/2003
End date DD/MM/YY**	20/10/2003

Figure 2: Screenshot of the Metadata Manager.

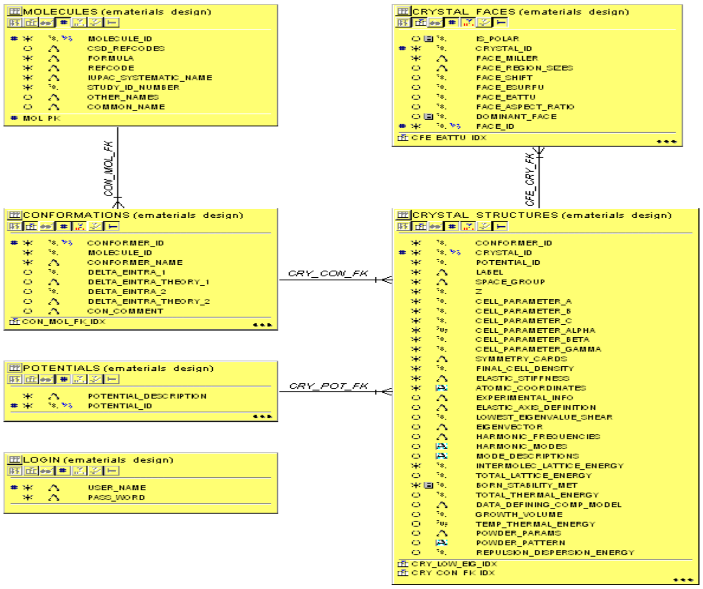


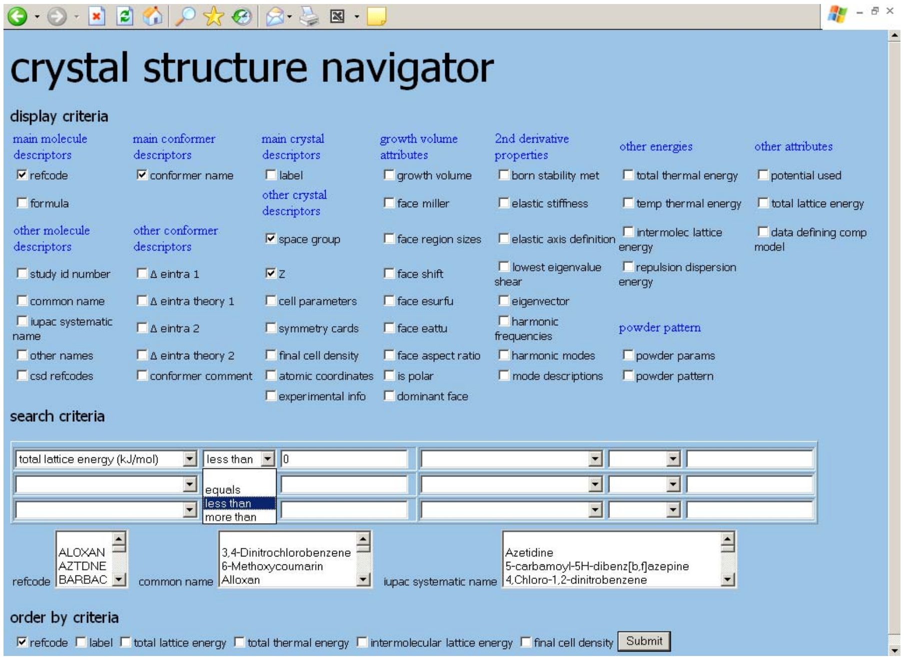
Figure 3: Crystal database schema

2.6. Populating the Database

An automated data ingestion agent, (written by Ian Brown at UCL), parses the different 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 with information from the study directory; otherwise new records are inserted into the database to hold this information.

2.7. 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 user defined queries against the calculated crystal database. 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. The flexibility of how to view and analyse the data is inherent in the application as the user has control over the query build, rather than only having a smaller set of predefined queries to choose from.



crystal structure navigator

display criteria

main molecule descriptors
- ☑ refcode
- □ formula

other molecule descriptors
- □ study id number
- □ common name
- □ iupac systematic name
- □ other names
- □ csd refcodes

main conformer descriptors
- ☑ conformer name

other conformer descriptors
- □ Δ eintra 1
- □ Δ eintra theory 1
- □ Δ eintra 2
- □ Δ eintra theory 2
- □ conformer comment

main crystal descriptors
- □ label

other crystal descriptors
- ☑ space group
- ☑ Z
- □ cell parameters
- □ symmetry cards
- □ final cell density
- □ atomic coordinates
- □ experimental info

growth volume attributes
- □ growth volume
- □ face miller
- □ face region sizes
- □ face shift
- □ face esurfu
- □ face eattu
- □ face aspect ratio
- □ is polar
- □ dominant face

2nd derivative properties
- □ born stability met
- □ elastic stiffness
- □ elastic axis definition
- □ lowest eigenvalue shear
- □ eigenvector
- □ harmonic frequencies
- □ harmonic modes
- □ mode descriptions

other energies
- □ total thermal energy
- □ temp thermal energy
- □ intermolec lattice energy
- □ repulsion dispersion energy

powder pattern
- □ powder params
- □ powder pattern

other attributes
- □ potential used
- □ total lattice energy
- □ data defining comp model

search criteria

total lattice energy (kJ/mol) | less than | 0

equals
less than
more than

refcode: ALOXAN, AZTDNE, BARBAC

common name: 3,4-Dinitrochlorobenzene, 6-Methoxycoumarin, Alloxan

iupac systematic name: Azetidine, 5-carbamoyl-5H-dibenz[b,f]azepine, 4,Chloro-1,2-dinitrobenzene

order by criteria

☑ refcode □ label □ total lattice energy □ total thermal energy □ intermolecular lattice energy □ final cell density

Figure 4: Crystal Navigator selection screen

From the display criteria section of the web page users can choose which combination of any of the forty-nine 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'. As such users have useful access to their data and can quickly and easily retrieve all the hypothetical crystal structures with the lowest binding energy, these crystals therefore being the most likely to be experimentally feasible, and the best candidates to try and create in the laboratory.

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 to them from 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 then download the results into a spreadsheet if they so wish. This application is now being used for the primary scientific analysis and for publications.

3. Conclusion

The project has supplied an effective infrastructure for the sharing of data and the provided suite of software is used daily to aid the project's collaboration.

Currently the applications are a little disjointed since they were rolled out in phases. Initially, the files storage was addressed along with metadata capture via the SRB and Metadata Editor. This enabled a catlogue to be created of the very many files that they held and allowed them to find and share their data via the Data Portal in a limited way. Later the extraction of data from the files themselves was enabled via the Data Ingestion Agent, the Computed Crystal Structure Database and the Crystal Navigator allowed the scientists to create much more complex queries and to truly study relationships within the results.

In future we plan to provide a more seamless interface between all the modules so that a search of the Data Portal would link not only to the files, but also to the crystal data in the database. This would provide a much richer interface for sharing data in the community.

The scientists also use other number of simulation codes on the same data further along the workflow. We would like to extract data from those files also and extend the crystal structure database to add pictures and visualizations of crystal structures. Furthermore it would be interesting to perform some matching against actual polymorphs found experimentally and store those parameters and pictures.

The use of metadata facilitates data browsing and discovery by others and so makes the sharing of data easier. However, whilst the metadata annotations encourages and enables knowledge sharing it is time-consuming and prone to error for the scientist to manually add metadata to files or collections of files. The RCommands, implemented by one of the co-authors of the paper, Dr. Richard Tyer, have been written to perform automatic metadata capture, and provide an automated version of the functionality of the Metadata Manager. In the future these scriptable commands could be used to insert and update metadata holdings.

References

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

[2] L. Blanshard, R.P. Tyer & K. Kleese-van-Dam eMaterials: Integrating Grid Computation and Data Management Services, *Proc. UK e-Science All Hands Meeting*, Nottingham, UK, 2004.

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

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

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

[6] L. Blanshard, K. Kleese van Dam, C.R.A. Catlow, S.L. Price, Simulation of Complex Materials: Database Design for Metadata, *Proc. UK e-Science AHM*, Nottingham, UK, 2003

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

[8] SRB Home Page http://www.sdsc.edu/srb

[9] G. Drinkwater, K. Kleese van Dam, A. Manandhar, S. Sufi, L. Blanshard, Data Management with the CCLRC Data Portal, *Proc. ICPDPTA*, Las Vegas, USA, 2004.

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

[11] D.J. Willock et al, *DMAREL*, J. Comput. Chem, 1995, 16, 628.

[12] W. Emmerich et al, Increasing the Scope for Polymorph Prediction using e-Science, *Proc. All Hands e-Science Conference*, Nottingham, UK, 2004