

The 'end to end' crystallographic experiment in an e-Science environment: From conception to publication.

S. Coles^a, J. Frey^a, M. Hursthouse^a, M. Light^a, L. Carr^b, D. DeRoure^b, C. Gutteridge^b, H. Mills^b,
K. Meacham^c, M. Surridge^c, L. Lyon^d, R. Heery^d, M. Duke^d, M. Day^d.

^aSchool of Chemistry, University of Southampton, Southampton, UK, ^bSchool of Electronics and Computer Science, University of Southampton, Southampton, UK, ^cIT Innovation Centre, University of Southampton, Southampton, UK, ^dUKOLN, University of Bath, Bath, UK.

Abstract

Recent developments at the UK National Crystallography Service (NCS), in collaboration with the CombeChem eScience testbed and the eBank-UK projects, have been aimed at developing an eScience infrastructure to facilitate the crystallographic experiment from end to end. A seamless distributed computing approach is shown to be able to transform a conventional but high throughput service, to enable access and secure remote operation with the visualisation of the diffraction experiment, through the data workup and analysis to the dissemination and further use of the resulting structural data. Access to use the NCS facilities and expertise and a mechanism to submit samples is granted through a secure Grid infrastructure. The user may then monitor and steer the data collection aspects of their experiments and results data staged to a securely accessible location. Publication of ALL the results data generated during the course of the experiment is then enabled by means of an Open Access Data Repository. This repository publicises its content through Open Archive Initiative (OAI) protocols, which enable harvester and aggregator services to make the data searchable and accessible via data portals.

1. Introduction

The recent advent of eScience is not only providing an infrastructure for research, but it is also ideal for the provision of services. The EPSRC National Crystallography Service (NCS) (<http://www.ncs.chem.soton.ac.uk>) is a UK national facility offering data collection or full structure solution services to the UK Chemistry community. NCS has a throughput in excess of 1000 samples a year in a laboratory environment that processes >2000 datasets per annum. As part of a UK National e-Science development program (<http://www.rcuk.ac.uk/escience/>) the CombeChem testbed project and the NCS developed a proof of concept demonstration outlining how the Grid could enable an e-Science enhancement for structural chemistry, which has subsequently been transformed into a functional service.

This paper describes the design of a Grid service for the NCS that aims to increase and enhance user interaction with experiments and provide efficient management of the resulting

data. The paper also outlines an institutional repository approach to the dissemination of the resulting experimental data. The ultimate aim is to use eScience to enable crystal structure results to be rapidly and efficiently generated, disseminated and reused.

The Grid can potentially provide many applications to Service Crystallography. The primary benefits would be in situations where the physical locations of the service and its user are distant. The Grid also provides a medium whereby an expert crystallographer (user) may control their own experiment remotely, to some extent independently of the service operator, or alternatively contribute specific knowledge of the sample to assist the on-site service operator. Once the data has been acquired the Grid can provide distributed software resources for the analysis of crystal structures, further data mining and 'value added' exercises (i.e. follow on data services after data collection and structure refinement). The Grid can similarly facilitate the efficient management of the data and rapid dissemination of results.

The implementation of such an infrastructure has a number of advantages over a conventional laboratory that would be of considerable worth to the service crystallographer (von Laszewski et al, 2000). A massively increased interaction between local experts and users would allow both chemists unskilled in the art of crystallography and 'trained' crystallographers alike to participate in, or conduct their own, experiments remotely. Enabling remote control for the expert user will assist in 'dark' laboratory instrument automation, allowing service operators to concentrate on other matters. Furthermore, automated data workup and structure solution and refinement software routines will increase the output of crystal structures and the resulting necessary rapid dissemination of results to the scientific community will be facilitated by the use of Grid technology.

1.1 The CombeChem project and NCS

The UK National Crystallography Service, funded by the EPSRC, is located in the School of Chemistry at the University of Southampton. The NCS maintains state-of-the-art X-ray diffraction experiment facilities, whereby external clients may submit chemical samples for processing, in order to derive their crystal structures. Different levels of service are provided, depending on the requirements or expertise of the user:

- Data collection & full structure determination (X-ray diffraction analysis of the sample, plus further analysis to derive the full crystal structure) or
- Data collection only: (for users with some expertise in structure solving; X-ray diffraction data is collected and basic processing is carried out, in order to generate a data file, which may subsequently be processed further by the user, to generate the full crystal structure).

The advanced facilities available to NCS allow them to process particularly hard-to-handle samples, such as extremely small or low melting point crystals, air-sensitive compounds and powders. NCS can perform screening of samples, to determine ones that may require more sensitive experiments, e.g. using the single

crystal diffraction facility at the Daresbury Synchrotron (Cernik, 1997).

The EPSRC Comb-e-Chem project aims to develop an e-Chemistry activity for the determination, assessment and utilisation of chemical structure information. The NCS is core to this activity, providing essential experimental services for the determination of crystal structures, and ideally placed to develop mechanisms to facilitate structure determination from experimental data. The NCS was established in 1982 and hence is an established service, although current working practices are very much "off-line". Currently there is no client interaction in the experimental process and instrument time is often wasted, attempting to analyse samples that are of poor quality or turn out to be uninteresting and will remain unpublished. It would be much more preferable if the client could be involved in the process (particularly in the early stages), to help the service operator in decision making. The client may even then help to steer the experiment, by adjusting parameters that the service operator would normally set. A Grid Service for the NCS is therefore a natural development of the service that is already available, empowering a client to steer their own experiment and gain faster access to their results data.

For the NCS Grid Service to work in a larger e-Science context, we must necessarily address issues such as:

- Authentication of clients
- Security of client data
- Provenance
- Interoperability with other services

1.2 The e-Bank-UK project and NCS

In crystallography in the 1960s, a doctoral student might have investigated three or so structures, now this number can be analysed in a single morning, yet the publishing protocols for reporting this work are essentially unchanged. Across the scientific domain, only a small percentage of data generated by many scientific experiments appears in, or is referenced by, the published literature (Hey, 2003). In addition, publication in the mainstream literature still

offers only *indirect* (and often expensive) access to this data. Moreover, the reuse of this data, in e.g. structural informatics studies, relies on mining as large a collection of crystallographic structure data as possible. As the access to and reuse of this data is dependent on it being released to the public domain in the traditional fashion, chemoinformatics and related fields are currently not as powerful as their potential suggests. As a consequence the user community is deprived of valuable information and funding bodies get a poor return on investments.

For the research chemist just 300,000 crystal structures are available in subject specific databases that have harvested their content from the published literature. It is estimated that 1.5 million structures have been determined in research laboratories worldwide and hence less than 20% of data generated in crystallographic work is reaching the public domain (Allen, 2002 & 2004). This shortfall is entirely due to current publication mechanisms. As high-throughput technologies, automation and e-science become embedded in chemical and crystallographic working routines, the publication bottleneck can only become more severe (Hursthouse, 2004). These facts are exemplified by the NCS statistics shown in figure 1, where over the last six years the number of crystal structures deposited in structural databases has remained roughly constant, despite a rising output, and comprises only ca 15% of the annual crystal structure turnover.

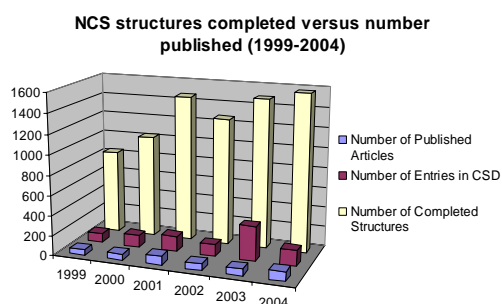


Figure 1. Publication and turnover statistics for NCS (1999-2004).

The eBank UK project has addressed this issue by establishing an institutional repository that supports, manages and disseminates metadata relating to crystal structure data. As part of the larger picture eBank is investigating the role of aggregator services in linking data-sets from Grid-enabled experiments to open data archives contained in digital repositories through to peer-reviewed articles as resources in portals.

2. Requirements, workflow and scenario for the NCS Grid Service

2.1 General requirements

General and overall objectives of the NCS Grid service may be considered to be:

- to allow remote users to interact with their experiments; The user should be able to submit a sample to the NCS, track the sample's progress through the system, and monitor the experiment(s) carried out on their sample. In addition there should be an ability to "steer" the experiment (but not drive it directly), either via an online conference with the service operator, or by direct adjustment of key experimental parameters.
- to provide users with better and faster access to experimental data; The Grid service should allow the user to access the raw data coming off the experiment (e.g. X-ray images), and gain faster access to the processed data.
- to provide a robust security infrastructure; Clients must be authorised to use the NCS Grid Service, by means of a Public Key Infrastructure (PKI). Clients must only be authorised to monitor (or steer) their own experiments, and access their own data with all other access restricted. All data transferred between the Grid Service and the client must be encrypted.
- to exploit this collaboration to improve NCS efficiency; With the user directly involved in the decision-making, efficiency is improved through reduction in wastage of diffractometer time (through unnecessary data collections or better understanding of sample quality), and allow service operators to concentrate on other tasks
- to be compatible with NCS operational processes; A Grid service must not impact heavily on current operational procedures
- to be a real (operational) service to which a user could easily subscribe; As opposed to previous demonstrator projects the NCS Grid service should be fully operational and should function robustly, reliably and securely
- that client software should be easy to set up and install; The Grid service has

to be implemented into an existing service with users with users possessing a wide range of computing expertise.

2.2 NCS Grid service scenario

Scenarios are presented assuming the processes of application, security, licensing and software installation have been undergone so that the detail of the experiment can be focussed on. Coloured text is employed so that the scenario can be understood from both the lab perspective and that of the user.

1. A new sample arrives at NCS and the administrator logs it in to the sample tracking database. An automatic email confirmation is received by the user.
2. At the end of the day NCS schedule samples for the next day. The user is emailed automatically and informed that they are second in the queue for the following day.
3. The following day the lab runs the first sample. The user is informed that their sample is next and an estimated start time given. When this time approaches the user logs on and waits for the experiment to start.
4. When the current sample is completed the scheduled sample status changes to running. The user may now initiate the control service software for that sample.
5. The lab starts the Automated Experiment Driver Software (AEDS), which automatically drives the experiment, but allows for input from the user). The lab portal detects the creation of a new directory and monitors for any files to appear. The control service reports to the user that the experiment has started.
6. The AEDS collects 4 scans to determine the quality of the crystal. As they are collected the user can view these scans.
7. On completion the AEDS analyses the scans, determines them to be suitable and gives the user the option to override. The user accepts.
8. The AEDS then proceeds to set parameters for a unit cell determination. The user views the parameters, may alter them if desired, and accepts. The AEDS collects the data (2-3 minutes), stages the raw file to the users directory and converts to JPG. The lab portal detects the JPG and transmits it to the user. The user views the images as they are generated. The AEDS calculates the unit cell on completion of the data collection and decides to collect the full dataset. The user

is presented with the unit cell calculation results, informed that the AEDS will continue and is asked whether they wish to override. The user accepts.

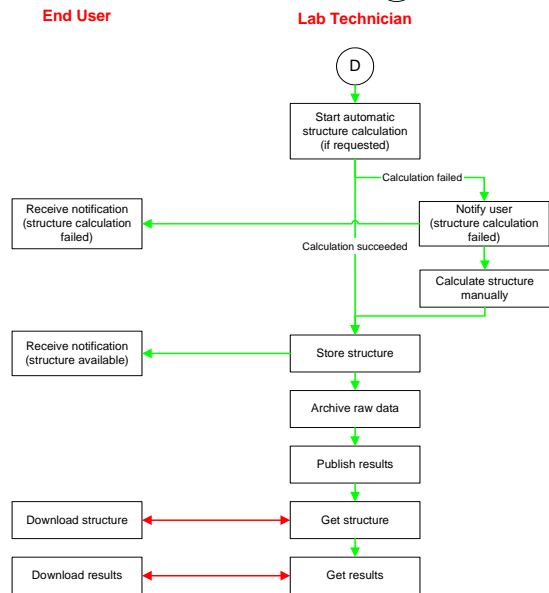
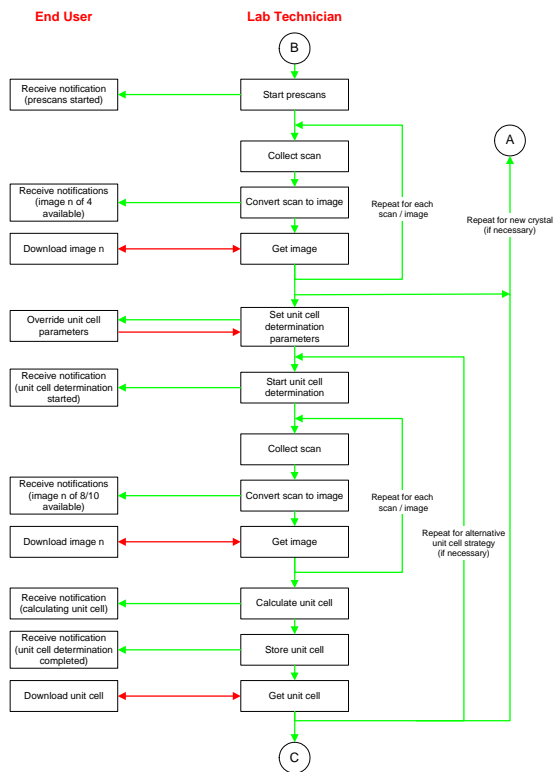
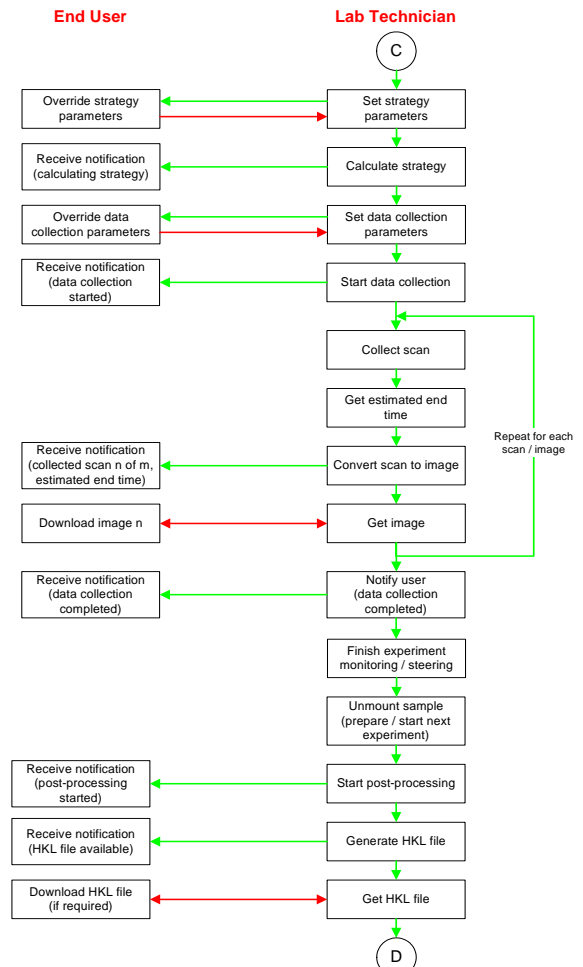
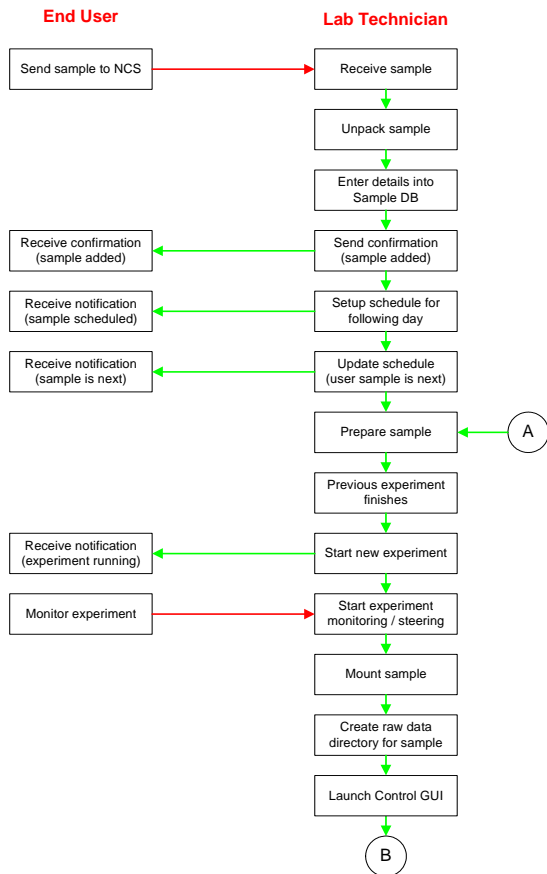
9. The AEDS calculates default parameters (time, distance, angle) and informs the user. The user is presented with data collection parameters that may be edited within limits and accepts.
10. The AEDS starts the full data collection and stages images as they are generated. The user views images as the instrument collects them and is dynamically informed of the completion time and number of images collected. The user may log out.
11. The AEDS emails the user on completion of the data collection and transforms the sample status to processing. The user waits for the lab to process the data. When a data file is generated the status of the sample is set to succeeded. The user may download the raw data and associated report.
12. The user decides to use the structure determination service and initiates it from the control service. Structure determination code is run and a result generated.
13. The service operator selects 'publish and archive results' and the structure data is deposited into the NCS structures database and an email sent to the user. The user receives an email reporting the successful experiment, logs onto the NCS Grid service and downloads their results data and report. The user is presented with options, such as 'visualise', or 'launch structure-property query' for further, value added calculations to be performed on the structure.

2.3 NCS Grid service workflow

The scenario outlined above was used to generate the workflow depicted in figures 2-5, which is sufficiently detailed to describe the entire Grid-based service crystallography experiment from the point of sample submission to final results download.

3 NCS Grid service architecture and implementation

The main components of the NCS Grid service architecture are outlined in Figure 6 and depict the firewall structure with respect to the laboratory, the user and the Grid service hardware.



Figures 2-5. Detailed workflow for a crystallographic experiment performed via the NCS Grid service

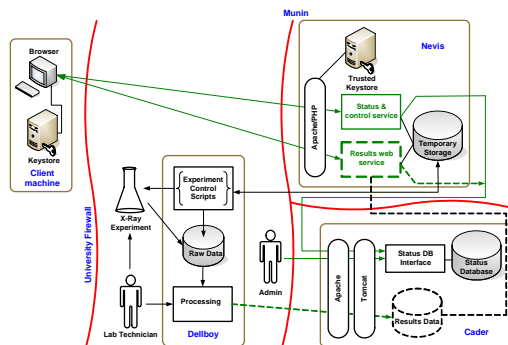


Figure 6. NCS Grid service architecture

3.1 Security and registration procedure

To apply for access to the NCS potential users must submit an online application form and accompany it with a case for support. The NCS RA verifies the identity provided by the applicant via independent means (e.g. telephone call using details from website) and forwards the application for review. If access is granted then the contact details provided in the application form are used as the Distinguished Name for that particular user when generating a security certificate. Service signed certificates are generated by the NCS RA. The certificate is written to a CD and sent to the users, independently verified, postal address and the NCS RA emails the passcode to the address in the DN. The NCS registration procedure is published at: (http://interact.xservice.soton.ac.uk/portal/cert_inst_guide.php)

3.2 Architecture

The X-ray diffractometer and associated hardware is located at NCS, within the University of Southampton network. The instrument is controlled by bespoke, manufacturer written, software manually driven by the service operator via a GUI. However, it is also possible to drive the diffractometer using command line calls, via an API. For the NCS Grid Service, scripts have been developed to drive the workflow normally carried out manually, which is essential as the experiment progresses, raw data is deposited into a unique working directory, which the user has no direct access to. The experimental data is made available by copying it to a secure location on the server. The control script also makes calls to the sample/status database, at various key points during the experiment, to change the status of the sample being analysed.

3.2 User interfaces

The client interface to both Control and Status services is deliberately lightweight with access through a standard web browser so that no client software installation is required. The Status database is the core of the system and the user functionality provided through the Java interface is shown in figure 7.

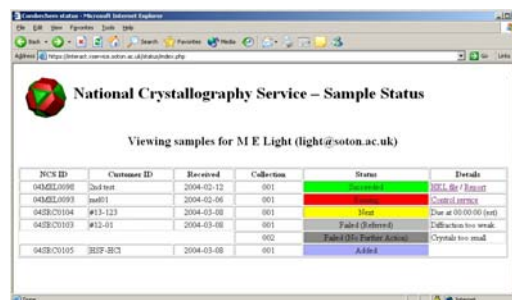


Figure 7. The Status service interface

The Status DB maintains tables of all submitted samples, details of users and mappings between and provides an externally visible web-based interface, as utilised by the Status and Control services. Each sample has an associated status; Added, Scheduled, Next, Running, Processing, Re-processing, Failed – no further action, Failed – referred and Succeeded and it is the Status of a sample that determines various authorizations within the system. The NCS Administrator uses various web pages to enter or modify sample data, rather than directly altering database entries. This service is written in PHP, which provides dynamic HTML for the user, whose access is via HTTPS and the PHP code determines the client's DN from their personal certificate. The DN is then used by the PHP code to query the Sample / Status database to obtain only the sample data owned by that DN.

The Control Service is also written in PHP, and provides a dynamic HTML interface (shown in figure 8) to the users' X-ray diffraction experiment. The Control service presents the client with certain key experimental parameters, which may be adjusted if necessary. The experiment control scripts provide suitable default values, and the user is given a time limit in which to enter new values, otherwise the experiment will proceed with the default values.

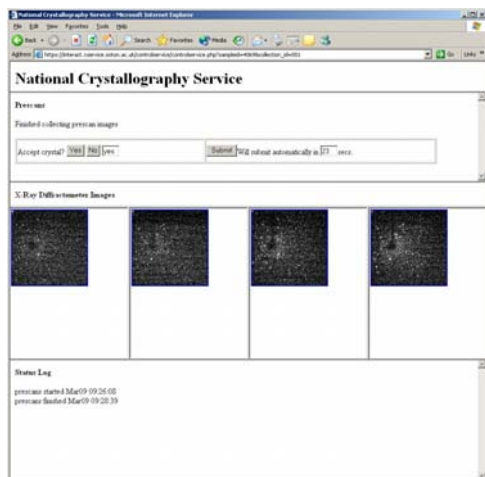


Figure 8. The Control service interface

4 Dissemination of crystal structure data via the Open Archive Initiative

The previous sections have been concerned with the acquisition of a set of raw experimental data. Now attention is turned to the management, publication and dissemination of the crystal structure of all data generated during the course of the experiment workup. The workflow for this is depicted in figure 9 (see p.8) with the corresponding digital files generated. For publication purposes the CIF file is currently perceived to be the final result of a crystallographic experiment, but in eBank value is added. For example CHECKCIF (<http://checkcif.iucr.org/>) is a web-based structure validation program, which produces an output file formatted as HTML and represents the *validation* stage of the process. In addition a CML file is generated (Chemical Markup Language) (Murray-Rust, 2001) that enables the exchange of this chemical structure information to be automatic in addition to being platform and software independent. The INChI identifier (<http://www.iupac.org/projects>) is included as a unique text representation of the molecule and has been shown to assist in the linking and aggregating processes (Coles, 2005).

The schema employed in the archive has the unique file extensions associated with the appropriate part of the process and therefore is able 'recognise' a particular file when it is presented. This enables a simple deposition process whereby the depositor supplies core bibliographic information and some chemical metadata, which is marked up as a combination of regular and qualified dublin core

(<http://dublincore.org>), along with a ZIP file containing the digital output from the experiment. In addition to making the files available, 'quality indicators' are extracted from a number of these files. This key information is presented alongside the files for download, the author input metadata and a rendered version of the CML file, which is made interactive through the use of an applet. When a stylesheet is applied to this data an entry in the archive is displayed as shown in Figure 10.

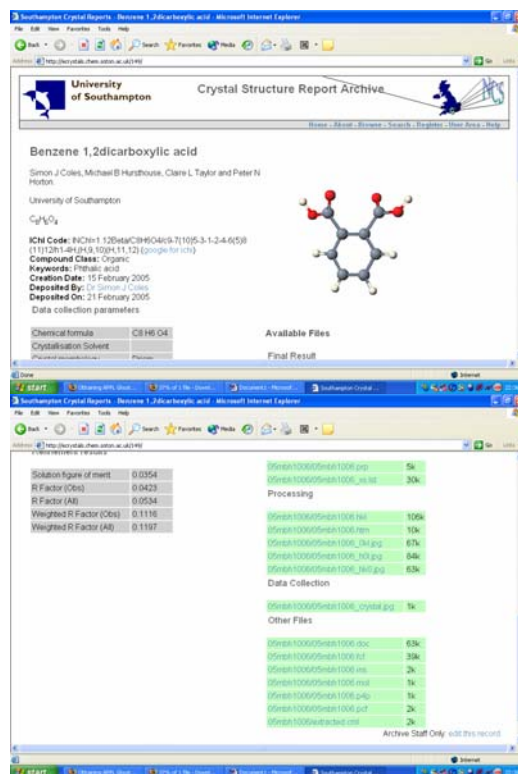


Figure 10. The local interface to the crystal structure repository

The repository publishes this metadata according to the Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH) (<http://www.openarchives.org/OAI/openarchiveprotocol.html>) the model for which has been extensively described as an output of the eBank-UK project (Heery, 2004, Lyon, 2004, and <http://www.ukoln.ac.uk/projects/ebank-uk/dissemination>). This metadata is intended to be used by service providers in the OAI-PMH model as a basis for building value-added services and a demonstrator for a crystal structure data aggregator service has been built (<http://www.ukoln.ac.uk/projects/ebank-uk/docs/technical/aggregator-description.html>). The findings of this demonstrator service have been the subject of recent work published by the

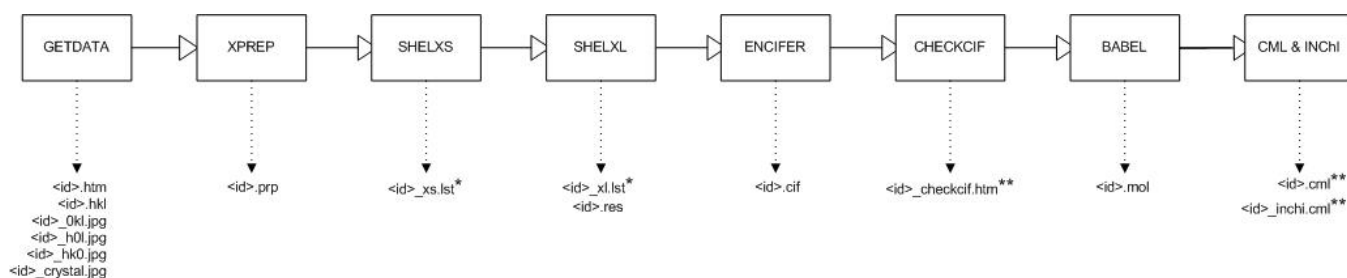


Figure 9. The crystal structure determination workflow

eBank-UK team (Duke, 2005) and will not be described further here.

5 Conclusions

A Grid infrastructure for conducting and monitoring service crystallography experiments and management, workup and publication of the subsequent results data has been outlined. This approach has been shown to facilitate the provision of an existing crystallography service whilst enhancing interaction and feedback with the experiment for the user. This approach, along with other recent technological advances, highlights a shortfall in the current publication and dissemination process, which has also been addressed. The operation of an OAI-PMH compliant crystal structure data repository has demonstrated the ability to open up access to research data by improving dissemination routes for the associated metadata.

Future developments are planned for the management of NCS Grid service workflows and data using the OAI repository and further investigations into dissemination and aggregation of crystal structure metadata are underway.

References

- Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst.*, B58, (2002), 380-388.
- Allen, F. H. High-throughput crystallography: the challenge of publishing, storing and using the results. *Crystallography Reviews*, 10 (2004), 3-15.
- Coles, S.J., Day, N.E., Murray-Rust, P., Rzepa, H.S., Zhang, Y., *Org. Biomol. Chem.*, 2005, (10), 1832-1834. DOI: [10.1039/b502828k](https://doi.org/10.1039/b502828k)
- Cernik, R. J., Clegg, W., Catlow, C. R. A., Bushnell-Wye, G., Flaherty, J. V., Greaves, G. N., Burrows, I., Taylor, D. J., Teat, S. J. & Hamichi, M. (1997). *J. Synchrotron Rad.* **4**, 279-286.
- Duke, M., Day, M., Heery, R., Carr, L.A., and Coles, S.J. Enhancing Access to Research Data: the

Challenge of Crystallography, JCDL 2005, Digital Libraries: Cyberinfrastructure for Research and Education, Denver, Colorado, USA June 7-11, 2005.

Hey, T. and Trethethen, A. The data deluge: an e-science perspective. In Berman, F., Fox, G. and Hey, A. J. G., eds., *Grid computing: making the global infrastructure a reality*. Wiley, Chichester, 2003, 809-824.

Heery, R., Duke, M., Day, M., Lyon, L., Hursthouse, M. B., Frey, J. G., Coles, S. J., Gutteridge, C. and Carr, L. A. Integrating research data into the publication workflow: the eBank UK experience. PV-2004: Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data (Frascati, Italy, October 5-7 2004).

Hursthouse, M. B. High-throughput chemical crystallography (HTCC): meeting and greeting the combichem challenge. *Crystallography Reviews*, 10, (2004), 85-96.

Lyon, L., Heery, R., Duke, M., Coles, S., Frey, J., Hursthouse, M., Carr, L. and Gutteridge, C. eBank UK: linking research data, scholarly communication and learning. Third UK e-Science Programme All Hands Meeting (AHM 2004) (Nottingham, UK, August 31 - September 3 2004).

Murray-Rust, P., Rzepa, H.S. and Wright, M. , Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content, *New J. Chem.*, 2001, 618-634.

von Laszewski, G., Westbrook, M., Foster, I., Westbrook, E. and Barnes, C. (2000), *Cluster Computing*, 3(3), 187-199.