

# An e-Science environment for service crystallography - from submission to dissemination.

*Simon J. Coles<sup>\*a</sup>, Jeremy G. Frey<sup>a</sup>, Michel B. Hursthouse<sup>a</sup>, Mark E. Light<sup>a</sup>, Andrew J. Milsted<sup>a</sup>, Leslie A. Carr<sup>b</sup>, David DeRoure<sup>b</sup>, Christopher J. Gutteridge<sup>b</sup>, Hugo R. Mills<sup>b</sup>, Ken E. Meacham<sup>c</sup>, Michael Surridge<sup>c</sup>, Elizabeth Lyon<sup>d</sup>, Rachel Heery<sup>d</sup>, Monica Duke<sup>d</sup>, Michael Day<sup>d</sup>.*

<sup>a</sup>School of Chemistry, University of Southampton, Southampton, SO17 1BJ, UK, <sup>b</sup>School of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK, <sup>c</sup>IT Innovation Centre, University of Southampton, Southampton, SO16 7NP, UK, <sup>d</sup>United Kingdom Office for Library Networking (UKOLN), University of Bath, Bath, BA2 7AY, UK.

s.j.coles@soton.ac.uk

## RECEIVED DATE

## Abstract

The UK National Crystallography Service (NCS) has developed a prototype e-Science infrastructure for the provision of a small molecule crystallography service from sample receipt to results dissemination. This paper outlines the two strands this service, which a) enable a user to contribute in the conduction of an experiment and b) provides an effective route for the archival and dissemination of the arising results. Access to use the NCS facilities and expertise and a mechanism to submit samples is granted through a secure Grid infrastructure, which seamlessly provides instantaneous feedback and the ability to remotely monitor and guide diffraction experiments and stage the diffraction

data to a securely accessible location. Publication of all the data and results generated during the course of the experiment, from processed data to analysed structures is then enabled by means of an open access data repository. The repository publishes its content through established digital libraries protocols, which enable harvester and aggregator services to make the data searchable and accessible.

## **Keywords**

Small Molecule Service Crystallography, Crystal Structure Datasets, Grid Computing, e-Science, Data Grid, Grid Middleware, Institutional Data Repository, Open Archives Initiative, Open Archives Initiative Protocol for Metadata Harvesting, Metadata Harvesting, Data Aggregation Service, Knowledge Management.

## **1. Introduction**

e-Science is computationally intensive science that is conducted in highly distributed network environments or uses large datasets that require Grid computing. The recent advent of e-Science is not only providing a modern IT aware infrastructure for research, but it is also ideal for the provision of services<sup>1</sup>, both physical and computational. There is considerable current interest in the concept of using e-Science and the Grid (the distributed computational infrastructure on which e-Science is conducted) to enable remote interaction with national and international instrument-based facilities. It is worth noting here that e-Science is a European term and a number of US projects at supercomputing centres use the term cyberinfrastructure to describe this type of work. Several major projects are in progress in this area, with a number that bear a particular relevance to the chemistry subject area. These include; ‘Common Instrument Middleware Architecture’<sup>2</sup> (CIMA), which is placing emphasis on supporting a variety of instrument and controller types; ‘SpectroGrid’<sup>3</sup>, which provides access to Nuclear Magnetic Resonance (NMR) instruments; ‘Remote Microscopy’<sup>4</sup>, which is concerned with access to the Imaging Technology Group facilities at the University of Illinois at Urbana-Champaign; ‘Virtual Laboratory’<sup>5</sup> (VLab), which

is concerned with the provision of the Nuclear Magnetic Resonance facilities at Poznan Institute of Bioorganic Chemistry. Additionally a consortium of Primarily Undergraduate Institutions in the USA (STaRBURSTT) are collaborating to build an infrastructure to provide access to scientific instruments for learning and training purposes<sup>6</sup>.

e-Science can potentially provide many applications for service crystallography<sup>7</sup> and in the current environment there is now a requirement for the rapid turnaround of analytical results as high throughput methods are being used. The Grid provides a infrastructure whereby an expert crystallographer (user) may control their experiment from a remote location, to some extent independently of the service operator, or alternatively contribute specific knowledge of the sample to assist the on-site service operator. In addition, automated data workup and structure solution and refinement software routines would facilitate the generation of crystal structures<sup>8</sup>. For example, once the data has been acquired the Grid can seamlessly provide access to 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). This software may be made available directly by the service provider, or access to software resources elsewhere may be negotiated through the service provider. The Grid can similarly facilitate the efficient archiving and management of the data and rapid dissemination of results with associated provenance details. Thirty years ago a research student would present about five crystal structures as their PhD thesis, however with modern technologies and good crystals this can now be achieved in the timespan of a single morning.

This increase in pace of generation further exacerbates a problem in the communication of the results. Across the whole scientific domain it is widely recognised that only a fraction of the data generated by scientific experiments appears in, or is referenced by, the published literature<sup>9</sup>. In addition, publication in the mainstream literature still offers only indirect (and often subscription controlled) 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 limited release to the public domain via traditional publishing procedures, chemoinformatics and related fields are currently not as powerful as their potential suggests. As a consequence the user community is deprived of valuable information.

For the academic research chemist approximately 500,000 small molecule (or small unit cell) crystal structures are available in subject specific databases that have harvested their content from the published literature, e.g. Crystal Structure Database (CSD)<sup>10</sup>, Inorganic Crystal Structure Database (ICSD)<sup>11</sup> and Metallic Structures Database (CRYSMET)<sup>12</sup>. It is estimated that 1.5 million small molecule structures have been determined in research laboratories worldwide<sup>13</sup> and hence a considerable proportion of the data generated in crystallographic work are not reaching the public domain<sup>14</sup>. This shortfall in data dissemination can in part be attributed 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<sup>3</sup>. These facts are exemplified by the statistics shown in Figure 1, which depicts the number of crystal structures produced by NCS alongside the arising number of CSD entries for the period 1999-2004. These figures exclude the 'data collection only' service provided by the NCS (which comprises about 60% of its output), local or collaborative research work and provision of a local service (which amounts to approximately the same output as the NCS). There are a number of reasons why a crystal structure might not be published through conventional routes, such as the result is not deemed significant enough, the researcher does not have the time to write a publication, the result is not that which was expected or the study was performed for inclusion in a thesis or project report only. These facts indicate that there is a considerable shortfall between the global small molecule crystallography output and that reaching the public domain.

Insert Figure 1 here

The eBank-UK (<http://www.ukoln.ac.uk/projects/ebank-uk/>) project has addressed this issue by establishing an institutional repository that supports, manages and disseminates metadata relating to crystal structure data (i.e. all the files generated during a crystal structure determination). As part of the

larger landscape eBank-UK is investigating the role of aggregator services in linking data-sets from Grid-enabled experiments to open data archives and through to peer-reviewed articles. This process alters the traditional method of peer review by openly providing crystal structure data where the reader / user may directly check correctness and validity. Accordingly, the provenance trail must be preserved so that potential follow-on amendments to, or comments on, the data may be properly captured and recorded.

The UK Engineering and Physical Sciences Research Council (EPSRC) funded National Crystallography Service (NCS) (<http://www.ncs.chem.soton.ac.uk>) is a national facility, operated out of the University of Southampton chemical crystallography group. The NCS offers either data collection or full structure solution services on small molecule systems, to the UK chemistry and crystallography communities. The NCS has a throughput of approximately 1000 samples a year in a world class laboratory (molybdenum rotating anode with focussing mirrors and CCD area detector and dedicated expert personnel) that processes approximately 2000 datasets per annum. As part of a UK National e-Science development program (<http://www.rcuk.ac.uk/escience/>) the CombeChem testbed project (<http://www.combechem.org>) and the NCS developed a proof of concept demonstrator<sup>15</sup> outlining how the Grid could enable an e-Science environment for structural chemistry, which has subsequently been transformed into a functional service.

The scope of this paper is to outline the design of a Grid service for the NCS that employs e-Science methodologies to enhance user interaction with experiments and provide efficient management, archival and dissemination of the resulting data. This service aims to act as a prototype for others and a full technical description of the two major components of this service (experiment interaction and results archival/dissemination) is in preparation for publication elsewhere.

## 2. Design of the NCS Grid Service

### 2.1 Requirements of a service

For a Grid Service providing experimental data capture, workup and publication facilities to be compliant with the larger e-Science context a number of issues must be addressed. These broadly include:

- Authentication of users; a client attempting to access the service must be authenticated as a bona fide current user.
- Security and integrity of user and service data; users must be authenticated to use the NCS Grid Service, by means of a Public Key Infrastructure (PKI). Clients must only be authorised to monitor (or guide) 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.
- Provenance tracking of data; data must be time stamped and associated with a particular project and user as it is created in the laboratory so that access is facilitated all the way down the data and analysis chains.
- Interoperability with other services; the design should employ current protocols and standards so that the service is compatible with related services using the same architecture.

Further aspects, that are specific to designing a remote interaction service for crystallographic experiments and placing it on the public network as part of the Grid, must also be considered. These are outlined below:

- Enabling remote users to interact with their experiments; the user must 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 guide the experiment, either via an online conference with the service operator, or by direct selection of key experimental parameters.

- Provide a real time notification service; a user must be made aware that an experiment has started and be able to monitor its progress through the system.
- Providing users with better and faster access to experimental data; the Grid service should allow the user to access the raw data as they are produced during the experiment (e.g. X-ray diffraction images), and gain faster access to the processed data.
- Exploitation of the enhanced collaboration to improve NCS efficiency; when a user is 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).
- Provision of a robust and operational service to which a user could easily subscribe, so that the system has a high degree of sustainability after the lifetime of a research project.
- Any user software must be simple to install as the Grid service is to be implemented into an existing service with users possessing a wide range of computing expertise. Usability is a crucial aspect of designing a service and must be taken into account at an early stage so that the software supports the user, rather than a user having to adopt the mentality of the software writer.

## **2.2 Analysis of the existing NCS workflow**

A first step towards designing such a complex system was the identification of the sequence of individual processes taken by users, service operators and samples, from an initial application to use the service to the final dissemination and further use of a crystal structure. All major activities, interactions and dependencies between the parties involved (both human and software components), may then be

described as a workflow, from which an architecture that would accommodate all the processes could be designed.

The workflow for a typical service crystallography experiment is quite complex when considered at this level of granularity. A typical Grid, or web, service would only involve computing components (e.g. calculations, data retrieval services), hence the workflow involving these services is fairly trivial to derive and can be automated by an appropriate workflow engine. However, the service crystallography workflow also includes many manual operations, e.g. sending a sample to the service or mounting a sample on a diffractometer. The derived workflow describing a Grid service for remote interaction with crystallographic experiments includes all possible processes, whether manual or automatic, from sending a sample, to downloading results at the conclusion of an experiment. This workflow is presented diagrammatically and deposited as supplemental information. The diagrams separate the end user, i.e. the person submitting a sample to the service, and the laboratory technician, i.e. the service operator running the experiment. In practice, there may be different people with a variety of roles, and some activities associated with the laboratory technician are now performed automatically by software components, but these must all be described by the workflow for a fully integrated service to be designed. In the diagrams, boxes represent activities which are linked in a sequence by arrows, where red arrows are initiated by the end user and green arrows are either initiated by the laboratory technician or automatically as part of a software process.

It is evident from this workflow that the NCS Grid Service, as shown by other scientific instruments on the Grid<sup>2,3,4,5</sup>, is server-driven as opposed to purely computational Grid services that are generally orchestrated by the user.

### **2.3 NCS Grid service architecture**



The architecture design for the Grid service is derived from the workflow, but must take into account other aspects, such as authentication, security and authorisation. The principal requirements in this respect are to:

- authenticate users to enforce access control rules
- protect the institution's network and contain any breaches of security inside the NCS system
- protect user data and control access to data based on user identity, but allowing a principal investigator to delegate access rights to a colleague
- synchronise access control with the laboratory process, i.e. instruments should be accessible by a user only when their own sample is being processed

These requirements, when combined with the derived workflow give rise to the security architecture for the NCS Grid Service, which is shown in Figure 2.

Insert Figure 2 here

The laboratory computer system which controls the diffractometer is shown in the centre of the diagram and is connected to the campus network and is therefore protected from the Internet by the University (or institutional) firewall. Two further firewalls have been implemented to secure the NCS Grid system, which are depicted on the right of the figure. The first firewall provides the NCS De-Militarised Zone (DMZ), whilst the second implements the NCS secure subnet which provides access to an experimental data staging system. Users can connect to the DMZ only. This is performed using HTTPS (Hyper Text Transport Protocol Secure), which requires a route through the campus network to accept incoming traffic on port 443 (as opposed to the normal port 80 used for standard Hyper Text Transfer Protocol - HTTP) and no special action is required by the user because any local firewall will be set up to allow outgoing HTTPS. The workflow management systems are managed by the DMZ computer systems and allow control of the instrumentation via a secure connection to the NCS subnet.

The NCS subnet is accessible only by means of Secure Shell (SSH) and only by the DMZ system and a subset of NCS computers in the laboratory (defined using IP addresses).

The workflow defined in the supplementary data gives rise to the design of a database which is core to the system and is capable of tracking a sample through the workflow. A sample is automatically given a different status in this database, according to its position in the workflow and each status has different authorisation conditions. The interplay between a generalised form of the workflow and the status of a sample is shown in Figure 3.

Insert Figure 3 here

The X-ray diffractometer is normally controlled by bespoke software manually driven by the service operator via a Graphical User Interface (GUI). However, it is also possible to drive the diffractometer using command line calls, via an Advanced Program Interface (API). So, for the NCS Grid Service, scripts have been developed to drive the workflow normally carried out manually, which is essential as the experiment must be run automatically. As the experiment progresses raw data are deposited into a unique working directory on the NCS subnet system, to which the user has no direct access. The necessary experimental data are made available to the user by copying to a secure location on the DMZ 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 Description of the NCS Grid service**

#### **3.1 Security and registration procedure**

The NCS Grid service security infrastructure is designed in accordance with a Public Key Infrastructure<sup>16</sup> (PKI) policy. This requires the validity of each party involved in a transaction to be

verified and authenticated using a system of X.509 digital certificates issued by a Certification Authority (CA) and a Registration Authority (RA). The issuing of certificates conforming to the X.509 specification requires adherence to a strictly-defined procedure<sup>17</sup>. Initially this was adopted, but credibility with the users required a slightly different approach. An alternative approach was devised<sup>18</sup> to avoid the requirement of users to install and use the relatively complex software used for the sign up and key management processes. The modified approach retains the software mechanisms, but handles the key generation centrally at the NCS. This deviates from a strict PKI in that user key-pairs as well as certificates are centrally generated, (i.e. by the NCS CA/RA), signed, and then securely transferred to the user, rather than relying on the user to perform the Certificate Signing Request (CSR) generation. This removes the risks of having users manage the key generation process using unfamiliar technology. It also allows NCS to leverage their existing trust relationships with users to manage the private key and certificate distribution as part of the user registration process. In this model the NCS takes over the RA function of validating the identity of each user before a key is issued. A close relationship, and therefore personal knowledge, is utilised to verify the identity of an applicant making a certificate request. A CSR is generated and validated by the NCS RA and a certificate generated and signed by the NCS CA in a similar fashion to a regular PKI policy. The current policy uses two routes (one non computer) to ensure an independent check of the identity of the requestor and also to transmit the signed certificate and its corresponding passcode. As user generation of private keys becomes more commonplace and the supporting software more user friendly, the NCS intends to adopt standard CA/RA CSR practice. Users are required to re-register annually to obtain an allocation and new certificates are issued accordingly. It is therefore possible to update the security infrastructure at the same time, should it be considered necessary to update or integrate with other schemes.

The NCS registration procedure and policy is published at:  
([http://interact.xservice.soton.ac.uk/portal/cert\\_inst\\_guide.php](http://interact.xservice.soton.ac.uk/portal/cert_inst_guide.php)).

The user access security requirements are handled using a Process-Based Access Security (PBAS) model, which allows each specific NCS process to validate user access according to defined requirements. The PKI mechanism is ideal for the PBAS model because each user certificate contains a Distinguished Name (DN) field which identifies the user, which enables NCS software to determine (for example) whether to allow data access if the user owns the sample that is being requested. Each specific NCS service can determine the access level allowed for a requesting individual depending both on the user certificate and the process actually running.

### **3.2 User interaction during the experiment**

The core of the NCS Grid service is the sample status database, which contains information on the position of the sample in the experimental workflow that may be updated by the system as processes are completed. A Status service written in Hypertext Preprocessor (PHP)<sup>19</sup> and running on the server visible to users, determines the DN of a user requesting access from their certificate and uses this to query the sample status database to obtain only the sample data owned by that DN. The statuses that a sample may be attributed with are outlined in Table 1 and the Status service as presented to the user is shown in Figure 4.

Insert Figure 4 here

Insert Table 1 here

#### **3.2.1 Preparation for the experiment**

On receipt of a sample the NCS administrator checks the sample details submitted by the user either via an on-line service or an accompanying paper form and enters them into the sample database. At this stage the sample is automatically given the *added* status. When the sample nears the top of the

queue a service operator changes the status to *scheduled*, at which point the user is automatically sent a notification email. When at the top of the queue the sample status is then set to *next* by a service operator, which alerts the user, via the status service, to the imminent start of the experiment.

### 3.2.2 Running an experiment

The x-ray diffractometer control software supplied with the instrument is written as a Python<sup>20</sup> library of command modules. In the case of the Grid enabled crystallography experiment, a scripted routine utilising the appropriate Python modules to enact the workflow allows automation of the data assessment and capture processes. Running in parallel to this, a Control service (written in PHP), provides a dynamic interface through which the user may easily interact with the experiment using a conventional internet browser. 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, after which the experiment will proceed with the default values.

At the point when the experiment is ready to start the service operator starts the experiment script, which automatically updates the status to *running* and provides a hyperlink in the status service that enables the user to participate with the experiment through the control service. The service operator may now leave the experiment for the user to monitor and/or guide.

The user is presented with initial diffraction images from the sample, which provide a comprehensive indication of its quality and suitability for data collection and is prompted for a decision to continue. Alternative methods for providing the quality of the diffraction pattern have been investigated, however a textual description cannot convey enough information and video conferencing methods are difficult to arrange and experience problems with band-width limitations and overcoming firewalls. The script then sets parameters for the initial unit cell determination, which the user may alter within defined boundaries. The raw data files generated during this short scan are sent to the sample

directory and converted to images (JPG format), which the user can view as they are being acquired. An example of the user interface to the control service is given in Figure 5. At the end of the scan the unit cell is determined and presented to the user, who may assess its validity and accept or reject. The script then calculates an appropriate data collection strategy and presents the parameters to the user, who may edit them within set limits. The data collection is then started and the user may monitor the raw data images as they are acquired.

Insert Figure 5 here.

On completion of the data collection the script sets the status to *processing*, extracts data from the raw image files and performs the necessary diffractometer dependent corrections using the standard manufacturer supplied software. The use of diffractometer software could potentially raise software license issues, but in this case the Grid service is built on an existing service for which site licenses had already been obtained. At this point, if the user has selected the data collection only service, the data files and an experiment report are uploaded to the sample database and the status service enables a hyperlink for the user to download. For this case the service operator would then set the sample status to *succeeded*. If the user has requested a full structure determination, this is performed by the service operator and when finished the sample may be set to *succeeded* as above and the user can download the completed structure data files and reports. The service personnel are then responsible for the transfer of the complete archived dataset to an archival service for long term storage and retrieval if necessary. The NCS grid service uses the Atlas Datastore<sup>21</sup>, based at the Rutherford Appleton Laboratory, UK. Approximately 1 Gb of data per day is transferred to the Atlas Datastore, who store the data with an off-site fire safe backup and migrate it on to new media as their service develops. Currently the data transfer is via File Transfer Protocol, but other front-ends to the Atlas Datastore are also provided, e.g. the Storage Resource Broker<sup>22</sup>.

A number of other statuses may be set by the service operator to provide feedback to the user when the data collection has finished. If there are problems with completing the processing or structure

determination and the service operator has to manually process the data the status is set to *reprocess* to inform the user. If the sample or data are found to be too poor to proceed with the experiment the status may be set to *failed (no further action)* or *failed (referred)*, depending on whether it is deemed worthwhile to send to follow-on services (e.g. the National Synchrotron Crystallography Service: [http://www.ncl.ac.uk/xraycry/srs\\_service.htm](http://www.ncl.ac.uk/xraycry/srs_service.htm)). There are current plans and activities focussed on linking the synchrotron component of the service to this grid system in the near future.

#### **4 Dissemination of crystal structure data via the Open Archive Initiative**

Technological advances in computing, instrument manufacture and now e-Science over the last three decades have led to an acceleration in the rate at which crystallographic data are generated. In addition, the general route for the publication of a crystal structure report is coupled with and often governed by the underlying chemistry and is therefore subject to the lengthy peer review process and tied to the timing of the publication as a whole. Mechanisms for the publication of a crystal structure report alone exist through the Acta Crystallographica series of journals (<http://www.iucr.org>), but these still remain fairly time-consuming procedures, as a full report must be written and subjected to peer review and editing.

One possible solution to this problem is to adopt the Open Archive Initiative (OAI) approach (<http://www.openarchives.org>) to the dissemination of information. To improve dissemination of published articles, this method allows researchers to share metadata describing papers that they make available in institutional or subject-based repositories<sup>23</sup>. Building on the OAI concept we have constructed an institutional repository that makes available all the raw, derived and results data from a crystallographic experiment (<http://ecrystals.chem.soton.ac.uk>), with little further researcher effort after the creation of a normal completed structure in a laboratory archive. Not only does this approach allow rapid release of crystal structure data into the public domain, but it can also provide mechanisms for value added services that allow rapid discovery of the data for further studies and reuse, whilst ownership of the data is retained by the creator. For publication without the peer review process it is

essential that all the necessary provenance information is provided so that users can access all the data generated during the experiment and then use this to self assess its validity and determine the exact processes used to derive the crystal structure report.

#### **4.1 The Open Access crystal structure report archive**

The archive is a highly structured database that adheres to a metadata schema which describes the key elements of a crystallographic dataset. Current details of this schema can be found at <http://www.ukoln.ac.uk/projects/ebank-uk/schemas/>. The schema requires information on bibliographic and chemical aspects of the dataset, such as chemical name, authors, affiliation etc, which must be associated with the dataset for validation and searching procedures. As standards must be adopted in order for the metadata in the archive to be compatible with that already accepted and available in the public domain a tool for aiding the deposition process has been built. This tool performs the necessary file format transformations and operations necessary for presentation of the dataset to the archive. The elements of the schema and a brief description of their purpose are given in Table 2.

Insert Table 2 here.

The metadata presented to the OAI interface falls into two categories. Institutional repositories are a mechanism for disseminating articles published in peer reviewed journals and a protocol, Dublin Core (DC),<sup>24</sup> has been developed for describing the bibliographic metadata that is made publicly available. The metadata that is disseminated by this protocol are; EPrintType, Subject, Title, Creator, Affiliation, Keywords, PublicationDate and Rights. The EPrintType is set to crystal structure, the Subject is chemistry and the Title is an International Union of Pure and Applied Chemistry (IUPAC) chemical name. It is important to note here that the generation of an IUPAC chemical name is not a trivial matter and a combination of chemical expertise and software routines are currently required to



perform this task. The recommendations in the guidelines and documentation for usage of this archive follow the current IUPAC conventions for generating a chemical name, as given in the Colour Books<sup>25</sup>.

However, a protocol for describing bibliographic information is insufficient for the dissemination of metadata regarding datasets. Fortunately the DC protocol contains a route around this problem by provision of Qualified Dublin Core, which allows for description of terms not contained within the kernel DC. The descriptions of terms falling into this category are made publicly available as an eXtensible Mark-up Language (XML) schema, so that any third party wishing to make use of this metadata may understand its meaning and incorporate it into their schema and processes. The metadata that are described in this manner are; ChemicalFormula, InChI, CompoundClass and AvailableData, and are included as identifiers to be utilised for subject specific services in the areas of discovery, harvesting, aggregation and linking.

The chemical formula is included, with guidelines for its composition, as a specific identifier to enable search and retrieval. The InChI (International Chemical Identifier)<sup>26</sup> is a relatively modern unique identifier which encodes molecular structure as a simple text string, with considerably more levels of description than any of its predecessors. In a recent development<sup>27</sup> the scope of InChI has been extended to include the phase of a compound and the crystalline phase descriptor may now be included to denote the fact that a particular InChI has been derived from crystal structure data. There are tools<sup>28</sup> available to generate InChI strings from common file formats (e.g. MOL, SDF), however this is a development project and there are still problems to overcome (e.g. description of polymers, complex organometallics, polymorphs) before an InChI can be used to describe all crystal structure data. An overview of the current capabilities of InChI may be found at <http://wwmm.ch.cam.ac.uk/inchifaq/>. As a result it is necessary to check the validity of a machine generated InChI. The archive deposition tool automatically generates an InChI string from a crystallographic dataset (via conversion to a MOL file) and then displays it in a rotatable, 2D form so that the depositor can check its integrity. As a text string InChI is easily machine readable and is included in an archive entry for the purposes of highly specific discovery and linking in the broader chemical literature. Initial studies<sup>29</sup> with linking data in different public

databases<sup>30</sup>, on the basis of an InChI have proven that indexing by the Google (<http://www.google.com>) search engine can give an exact match and may therefore potentially be used as a means of aggregating chemical information. The compound class element is for broad aggregation of datasets within the area of chemistry and is defined as organic, inorganic, bio-organic or organometallic. The 'available data' declares what categories of the experimental process have files associated with them and these are defined as stages thus; processing, solution, refinement, validation and final result and other files (where any files not recognised by the schema are placed).

On completion of the refinement of a crystal structure all the files generated during the process are assembled and deposited in the archive, a process that will be automated as part of future developments. The metadata to be associated with this dataset is generated at this point, either by manual entry through a deposition interface or by internal scripting routines in the archive software which extract information from the data files themselves. All the metadata are then automatically assembled into a structured report (see Figure 6) and an interactive rendering of a Chemical Markup Language<sup>31</sup> (CML) file added for visualisation purposes.

Insert Figure 6 here

For conventional publication purposes a crystal structure determination would normally terminate at the creation of a Crystallographic Information File (CIF)<sup>32</sup> and this file would be all that is required for submission to a journal. However, this archive enables publication of all the files generated during the experiment and moreover, during deposition a number of additional processes are performed which provide added value to the study and enable discovery and reuse of the data. These processes are seamlessly performed by uploading all the files up to and including the CIF to a toolbox on the archive server which can perform the necessary additional services required for a full archive entry. At this point validation of the structure is performed using the web service CHECKCIF<sup>33</sup>. The generation of the InChI and translation of the structure into CML format generates files for the final results stage which

are machine readable and therefore allow automatic processing of an entry by third parties. When deposited the new archive entry is queued to be further checked and signed off by an editor. A trained crystallographer would assume this editorial role and therefore provide further validation of the data prior to making it publicly available.

The funding councils in the United Kingdom have stated that '*the data underpinning the published results of publically-funded research should be made available as widely and rapidly as possible*'<sup>34</sup> which is also a similar stance to that adopted by the National Institutes of Health in the United States<sup>35</sup>. The approach outlined above provides a rapid and effective method of dissemination in accordance with the mandates of the funding bodies. In addition to the method described in this paper for addressing this issue there are other projects also making crystallographic data available in the public domain, most notably the ReciprocalNet<sup>36</sup> initiative to which 20 institutions around the world are contributing crystal structure data. The software described in this paper will be installed in a small number of UK institutions in early 2006 for testing purposes and will be made freely available after this exercise. This approach has led NCS to define a policy whereby all crystal structures determined will be made publicly available (unless specific reasons have been provided for withholding the data) on an open archive if the results have not been published within three years of the date of data acquisition. This policy ensures that a researcher has sufficient time to consider the results and prepare a publication (three years is deemed suitable as it is the timescale of a PhD studentship or a postdoctoral position), whilst enabling rapid dissemination if the result is destined not to be included in a traditional publication.

#### **4.2 Metadata harvesting and value added services**

When an archive entry is made public the metadata are presented to an interface with the internet in accordance with the Open Archive Initiative – Protocol for Metadata Harvesting (OAI-PMH)<sup>37</sup>. OAI-PMH is an accepted standard in the digital libraries community for the publication of metadata by

institutional repositories which enables the harvesting of this metadata. Institutional repositories and archives that expose their metadata for harvesting using the OAI-PMH provide baseline interoperability for metadata exchange and access to data, thus supporting the development of service providers that can add value. Although the provision of added value by service providers is not currently well developed a number of experimental services are being explored. The eBank UK project has developed a pilot aggregator service<sup>38</sup> that harvests metadata from the archive and from the literature and makes links between the two. The service is built on DC protocols and is therefore immediately capable of linking at the bibliographic level, but for linking on the chemical dataset level a different approach is required. The Dublin Core Metadata Initiative (DCMI) provides recommendations for including terms from vocabularies in the encoded XML, suggesting: *Encoding schemes should be implemented using the xsi:type attribute of the XML element for property.* As there are not as yet any designated names for chemistry vocabularies, for the purposes of building a working example the project defined some eBank terms as designators of the type of vocabulary being used to describe the molecule. Thus a chemical formula would be expressed in the metadata record as:

```
<dc:subject xsi:type="ebankterms:ChemicalFormula">C27H48</dc:subject>
```

In the longer term, it would be desirable if standardised methods were agreed within the chemistry community for defining the terms that designate a specific naming convention, using namespaces to support XML processing. The eBank terms are published at (<http://www.rdn.ac.uk/oai/ebank/20050617/ebankterms.xsd>), but at present are considered to be placeholders until official ones become available. Figure 7 depicts the representation and linking of resources in this pilot service.

Insert Figure 7 here

There are currently no journals publishing crystal structure reports that disseminate their content through OAI protocols. In order to provide proof of concept for the linking process the Rich Site

Summary (RSS) feed for the International Union of Crystallography's publications website was used to provide metadata and crystal structure reports published in these journals were then deposited in the archive, thus providing the aggregator service with two sources of information. Aggregation is performed on the following metadata; author, chemical name, chemical formula, compound class, keywords and publication date, thus providing a search and retrieval capability at a number of different chemical and bibliographic levels. The demonstrator system, along with searching guidelines may be viewed and used at the following address: <http://eprints-uk.rdn.ac.uk/ebank-demo/>.

## **5 Conclusions**

An e-Science infrastructure for conducting and monitoring small molecule crystallography experiments and management, workup and publication of the subsequent results data has been outlined and acts as an end-to-end prototype for the provision of such services. The resulting service 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.

## **Acknowledgements**

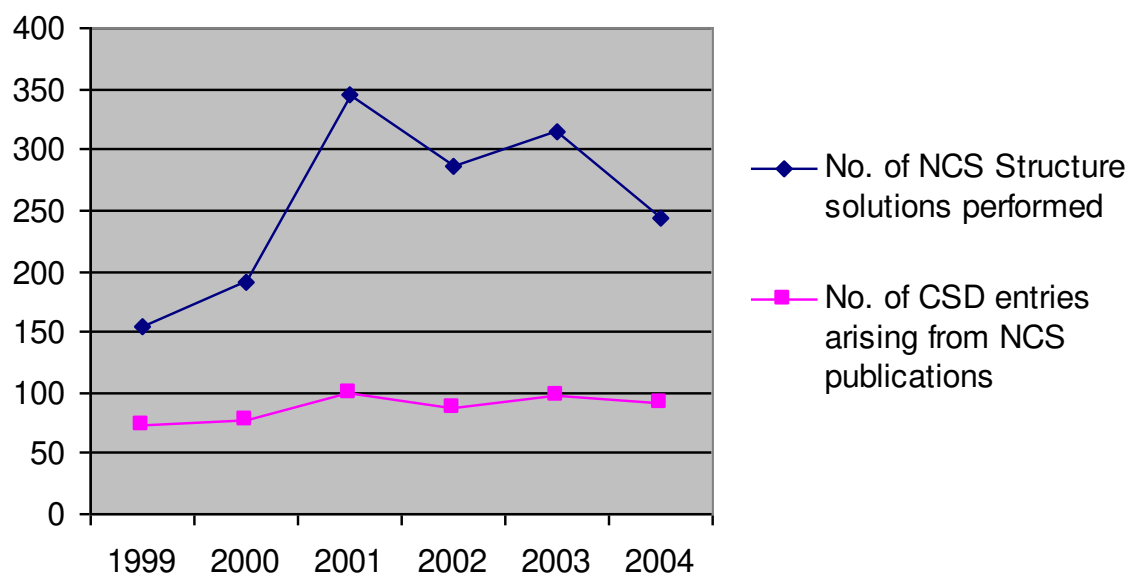
The authors acknowledge the provision of funding from the Engineering and Physical Sciences Research Council, EPSRC (CombeChem and NCS), the Department of Trade and Industry and the Joint Infrastructure Systems Committee, JISC (eBank). The authors would also like to thank the International Union of Crystallography for provision of old RSS feeds pertaining to articles with corresponding crystal structure data subsequently deposited in the archive.

## Supporting Information

Supplemental data deposited contains figures representing the workflow of a service crystallography experiment.

## Figures and captions

Figure 1. NCS structure determinations performed and the arising number of CSD entries (1999-2004).



† These are the statistics for the National Crystallography Service structure determination component only, which comprises approximately 40% of the data sets collected (the remaining 60% of data sets are collected as part of a 'data collection only' service). These statistics also do not include the research

interests, personal collaborations or local service provision performed by the Chemical Crystallography Group at the University of Southampton (which roughly equals the output of the NCS).

Figure 2. NCS Grid service architecture

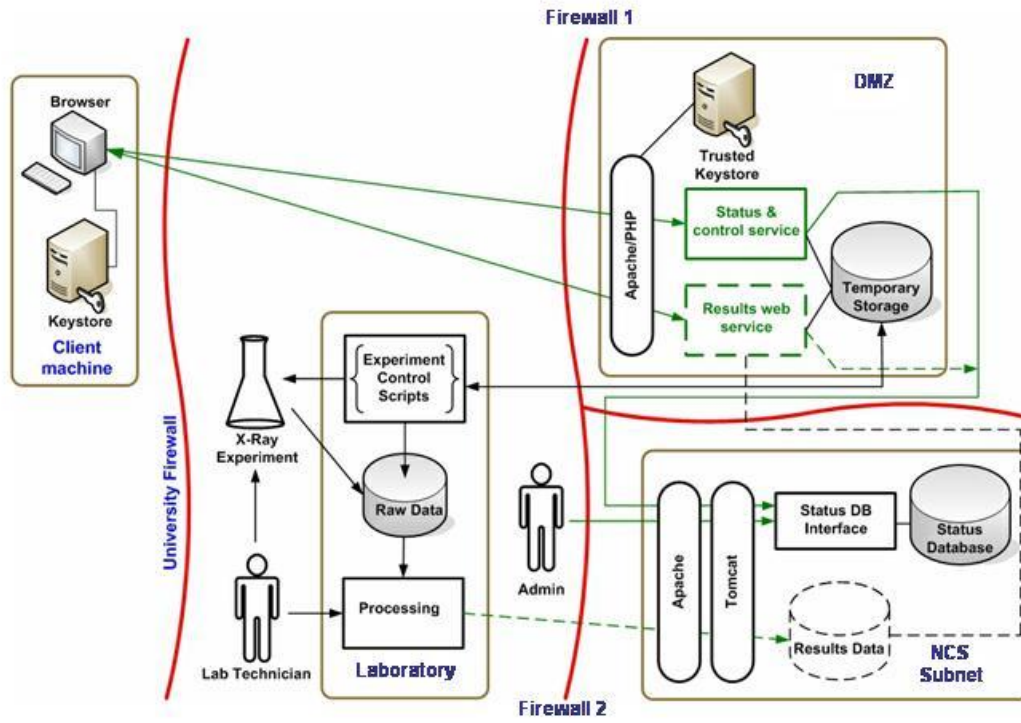






Figure 3. A generalised view of the interplay between the workflow and the status of a sample

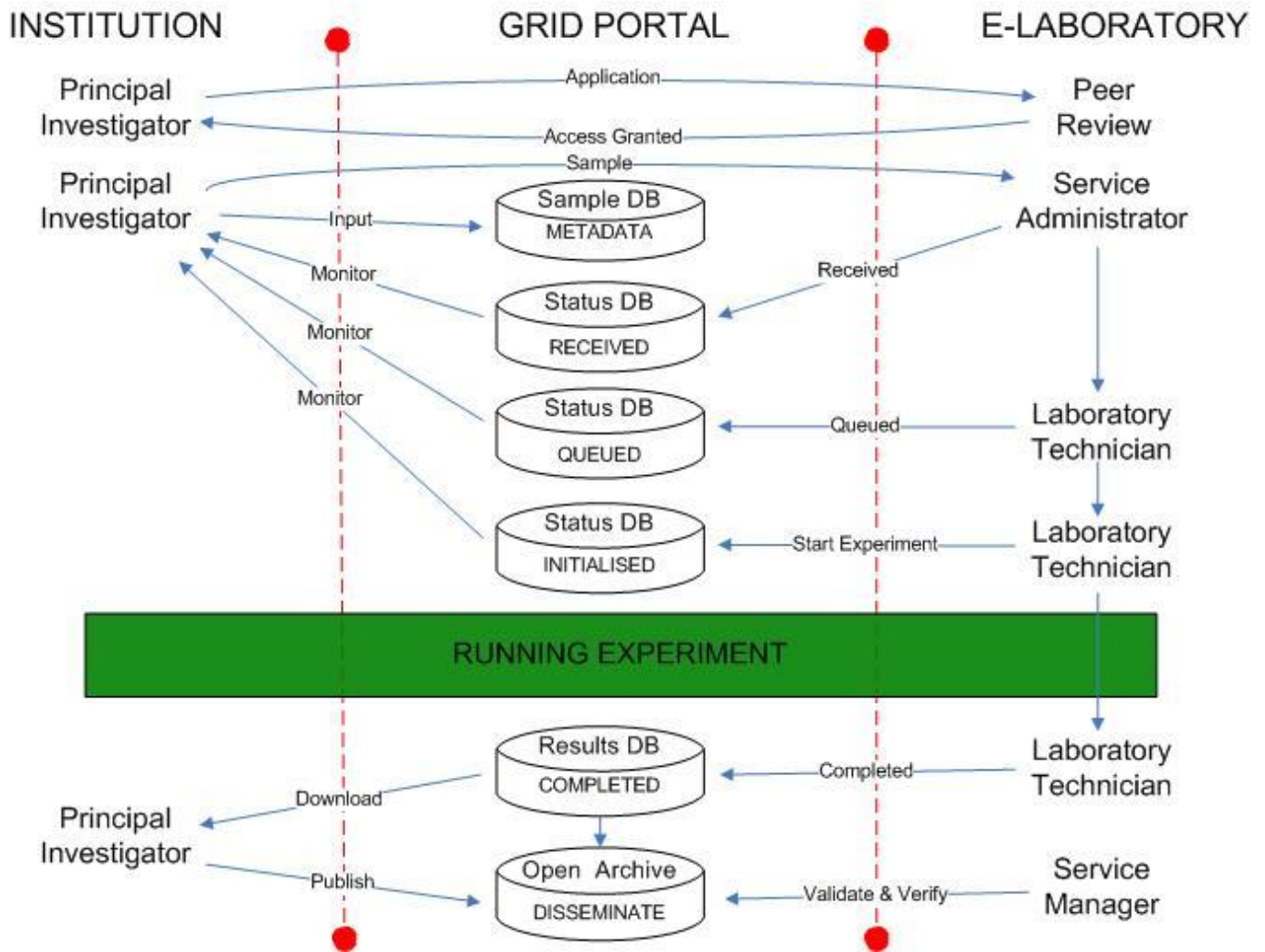




Figure 4. An example of the Status service interface presented to the user.

**National Crystallography Service – Sample Status**

Viewing samples for M E Light (light@soton.ac.uk)

NCS ID	Customer ID	Received	Collection	Status	Details
04MEL0098	2nd test	2004-02-12	001	Succeeded	<a href="#">HKL file / Report</a>
04MEL0093	me101	2004-02-06	001	Running	<a href="#">Control service</a>
04SRC0104	#13-123	2004-03-08	001	Next	Due at 00:00:00 (est)
04SRC0103	#12-01	2004-03-08	001	Failed (Referred)	Diffraction too weak
			002	Failed (No Further Action)	Crystals too small
04SRC0105	HSF-HCl	2004-03-08	001	Added	

Figure 5. An example of the user interface to the control service

National Crystallography Service - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address [https://interact.xservice.soton.ac.uk/controlservice/controlservice.php?sampleid=4069&collection\\_id=001](https://interact.xservice.soton.ac.uk/controlservice/controlservice.php?sampleid=4069&collection_id=001) Go Links »

## National Crystallography Service

### Data Collection

Crystal to detector dist / mm   Will submit automatically in  secs.

### X-Ray Diffractometer Images

### Status Log

```
prescans started Mar09 09:26:08
prescans finished Mar09 09:28:39
prescans accepted Mar09 09:29:16
unitcell started Mar09 09:29:47
unitcell finished Mar09 09:35:23
unitcell accepted Mar09 09:35:54
```

Figure 6. An archive entry for one dataset.

University of Southampton **Crystal Structure Report Archive**

Home About Browse Search Register User Area Help Single

**5- Cyano- 2- methyl- 4- phenyl- 1- (5,6,7- tris(acetoxy)- 2,10- dioxo- 3,9- dioxo- undeca- 4- yl)- 2- aza- 7- thiabicyclo[2.2.1]heptane- 3- one**

M. J. Arevalo, M. Avalos, R. Babiano, P. Cintas, M. B. Hursthouse, J. L. Jimenez, M. E. Light and J. C. Palacios.

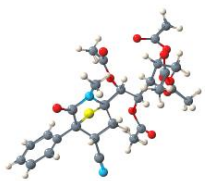
University of Southampton  
C28H32N2O11S

**InChI Code:** C28H32N2O11S,1H3-15(31)37-14H2-23H(38-16(2H3)32)24H(39-17(3H3)33)25H(40-18(4H3)34)26H(41-19(5H3)35)27-13H2-22H(7-29)28(42-27,21(36)30(27)6H3)20-11H-9H-8H-10H-12H-20 (google for ichi)

**Compound Class:** Organic

**Keywords:** Controlled Keywords UNSPECIFIED

**Creation Date:** 26 September 2001  
**Deposited By:** Susanne L. Huth  
**Deposited On:** 03 August 2004

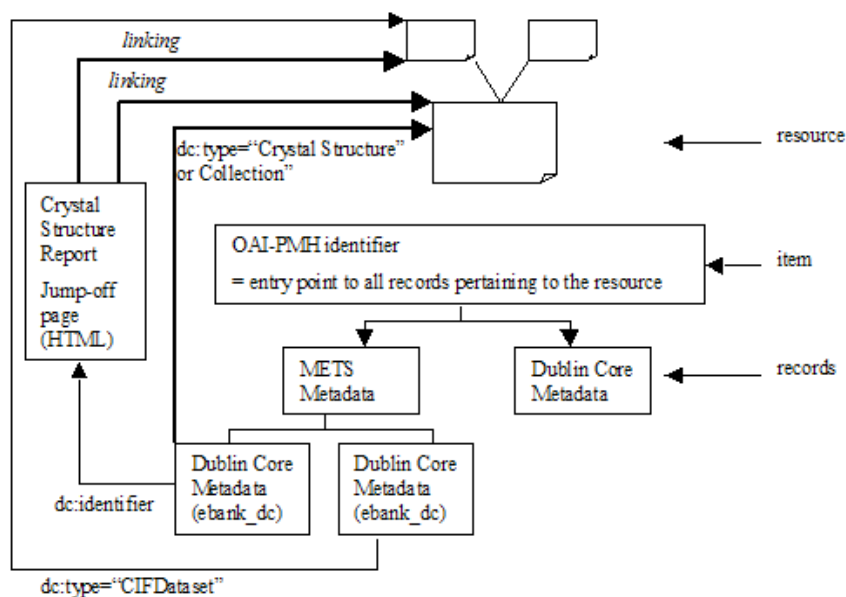


**Available Files**

**Final Result**

Data collection parameters		01esp301_data/01esp301.CIF	
Chemical formula	C28 H32 N2 O11 S	01esp301_data/01esp301.cml	11k
Crystallisation Solvent		01esp301_data/01esp301_inchi.cml	1k
Crystal morphology		<b>Validation</b>	
Crystal system	Orthorhombic	01esp301_data/01esp301_cif.html	15k
Space group symbol	P2(1)2(1)2(1)	<b>Refinement</b>	
Cell length a	10.9877(7)	01esp301_data/01esp301.res	10k
Cell length b	11.9703(8)	01esp301_data/01esp301_xl.lst	48k
Cell length c	22.4663(18)	<b>Solution</b>	
Cell angle alpha	90.00	01esp301_data/01esp301.PRP	6k
Cell angle beta	90.00	01esp301_data/01esp301_xs.lst	47k
Cell angle gamma	90.00	<b>Processing</b>	
Data collection temperature	120(2)	01esp301_data/01esp301.HKL	468k
		01esp301_data/01esp301.HTM	6k
<b>Refinement results</b>		<b>Other Files</b>	
Solution figure of merit	0.1386	01esp301_data/01esp301.DOC	290k
R Factor (Obs)	0.0848	01esp301_data/01esp301.mol	6k
R Factor (All)	0.3088	Archive Staff Only: edit this record	
Weighted R Factor (Obs)	0.1318		
Weighted R Factor (All)	0.1930		

Figure 7. Representation and linking of resources for the eBank aggregator service



## Tables and captions

Table 1. The allowed statuses for a sample

<b>ID</b>	<b>Status</b>	<b>Meaning</b>
1	Added	Sample has been added to the database.
2	Scheduled	X-ray experiment has been scheduled for this sample.
3	Next	Sample is next in experiment queue.
4	Running	Sample X-ray experiment is currently running.
5	Processing	Experiment post-processing underway for this sample.
6	Re-processing	Sample is being re-processed.
7	Failed – no further action	Experiment has failed for this sample. No further retries or referrals will take place.
8	Failed - referred	Experiment has failed for this sample. Sample has been referred (e.g. outsourced to a different experiment facility).
9	Succeeded	Experiment and post-processing has completed successfully.

Table 2. The metadata elements in the open archive schema.

<b>Metadata Element Name</b>	<b>Content Description</b>
EPrintType	Type of entry (e.g. crystal structure)
Subject	Subject discipline (e.g. crystallography, chemistry)
Title	IUPAC chemical name
Creator	Author(s)
Affiliation	Institution(s) of author(s)
Publisher	Publisher of a dataset (usually the institution)
ChemicalFormula	Formula of compound or moieties (according to IUPAC convention)

InChI	International Chemical Identifier (unique text identifier for a molecule)
CompoundClass	Chemical category (e.g. bio organic, inorganic)
Keywords	Selected keywords (provided as a limited ontology)
AvailableData	Stages of the experiment/determination for which datafiles are present (e.g. data collection, refinement, validation)
PublicationDate	Date when entry was made publicly available
Rights	Intellectual Property Rights exercised by the publishing institution

## References

- 
- (1) Hey, T; Trefethen, A.E., Cyberstructure for e-Science, *Science*, **2005**, *308*, 817-821 and references therein.
- (2) Bramley, R.; Chiu, K.; Huffman, J.C.; Huffman, K.; McMullen, D.F., Instruments and Sensors as Network Services: Making Instruments First Class Members of the Grid, *Indiana University CS Department Technical Report 588*, **2003**.
- (3) [http://nmr-rmn.nrc-cnrc.gc.ca/spectrogrid\\_e.html](http://nmr-rmn.nrc-cnrc.gc.ca/spectrogrid_e.html)
- (4) [http://www.itg.uiuc.edu/technology/remote\\_microscopy/](http://www.itg.uiuc.edu/technology/remote_microscopy/)
- (5) <http://www.terena.nl/library/tnc2004-proceedings/papers/meyer.pdf>
- (6) <http://www.as.yzu.edu/~adhunter/STaRBURSTT/index.html>
- (7) von Laszewski, G.; Westbrook, M.; Foster, I.; Westbrook, E.; Barnes, C., Using computational grid capabilities to enhance the capability of an X-ray source for structural biology, *Cluster Computing*, **2000**, *3*, 187-199.



- 
- (8) Hursthouse, M.B., High-throughput chemical crystallography (HTCC): meeting and greeting the combichem challenge. *Crystallography Reviews*, **2004**, *10*, 85-96.
- (9) Hey, T.; Trethethen, A.E., The data deluge: an e-science perspective. In Berman, F., Fox, G.; Hey, A. J. G., eds., *Grid computing: making the global infrastructure a reality*. Wiley, Chichester, **2003**, 809-824.
- (10) Allen, F.H., The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst.*, **2002**, *B58*, 380-388.
- (11) Belsky, A.M.; Hellenbrandt, M., V. L.; Karen, V.L.; Luksch, P., New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. *Acta Cryst.*, **2002**, *B58*, 364-369.
- (12) White, P.S.; Rodgers, J.R.; Le Page, Y. CRYSTMET: a database of the structures and powder patterns of metals and intermetallics. *Acta Cryst.*, **2002**, *B58*, 343-348.
- (13) Bond, A.D.; Davies, J.E., Data overload, *Chemistry in Britain*, **2003**, *39*, 44.
- (14) Allen, F.H., High-throughput crystallography: the challenge of publishing, storing and using the results. *Crystallography Reviews*, **2004**, *10*, 3-15.
- (15) Coles, S.J; Frey, J.G.; Hursthouse, M.B.; Light, M.E.; Meacham, K.E.; Marvin, D.J.; Surridge, M., ECSES -Examining crystal structures using e-Science: A demonstrator employing WEB and GRID services to enhance user participation in crystallographic experiments. *J. Appl.Cryst.*, **2005**, *38*, 819-826.
- (16) Nash, A.; Duane, W.; Joseph, C.; Brink, O.; Duane, B. PKI: implementing and managing E-security, 2001, New York: Osborne/McGraw-Hill.
- (17) Guida, R.; Stahl, R.; Blunt, T.; Secret, G.; Moorcones, J., Deploying and using public key technology, *IEEE Security and Privacy*, **2004**, *4*, 67-71.

---

(18) Bingham, A.; Coles, S.; Light, M.; Hursthouse, M.; Peppe, S.; Frey, J.; Surridge, M.; Meacham, K.; Taylor, S.; Mills, H.; Zaluska, E. Security experiences in Grid-enabling an existing national service, Conference submission to: eScience 2005, Melbourne, Australia.

(19) <http://www.php.net/>

(20) <http://www.python.org/>

(21) <http://www.e-science.clrc.ac.uk/web/services/datastore>

(22) <http://www.sdsc.edu/srb/>

(23) Lynch, C.A., Institutional repositories: essential infrastructure for scholarship in the digital age. *ARL Bimonthly Report*, **2003**, 226.

(24) <http://dublincore.org/documents/usageguide/bibliography.shtml> and references therein

(25) Blue Book (Guide): A Guide to IUPAC Nomenclature of Organic Compounds. Blackwell Scientific Publications, Oxford, 1993; Purple Book: IUPAC Compendium of Macromolecular Nomenclature. Second Edition, Blackwell Scientific Publications, Oxford, 1991; Red Book: IUPAC Nomenclature of Inorganic Chemistry. Third Edition, Blackwell Scientific Publications, Oxford, 1990; White Book: IUBMB Biochemical Nomenclature and Related Documents. Second Edition, Portland Press, London, 1992.

(26) Stein, S.E.; Heller, S.R.; Tchekhovski, D., An Open Standard for Chemical Structure Representation—The IUPAC Chemical Identifier, *Nimes International Chemical Information Conference Proceedings*, 2003, 131–143; Stein, S.E.; Heller, S.R.; Tchekhovskoi D.V., *Abstracts of Papers, 222nd ACS National Meeting*, Chicago, IL, **August 26–30 2001**, CINF-005.

- 
- (27) Brown, I.D.; Abrahams, S.C.; Berndt, M.; Faber, J.; Karen, V.L.; Motherwell, W.D.S.; Villars, P.; Westbrook, J.D.; McMahon, B. Report of the working group on crystal phase identifiers. *Acta Cryst.*, **2005**, *A61*, 575-580.
- (28) For example: OpenBabel (<http://openbabel.sourceforge.net/RELEASE.shtml>); Marvin (<http://www.chemaxon.com/marvin/>); ChemSketch (<http://www.acdlabs.com/download/chemsk.html>) and Open Source (<http://sourceforge.net/projects/inchi>).
- (29) Coles, S.J.; Day, N.E.; Murray-Rust, P.; Rzepa, H.S.; Zhang, Y., Enhancement of the chemical semantic web through the use of InChI identifiers. *Org. Biomol. Chem.*, **2005**, *10*, 1832-1834.
- (30) For example: <http://wwmm.ch.cam.ac.uk/data/kegg> and <http://www.epa.gov/nheerl/dsstox/DSSToxDatabases>
- (31) Murray-Rust, P.; Rzepa, H.S.; Wright, M. , Development of Chemical Markup Language (CML) as a System for Handling Complex Chemical Content, *New J. Chem.*, **2001**, 618-634.
- (32) Brown, I. D.; McMahon, B. CIF: the computer language of crystallography., *Acta Cryst.*, **2002**, *B58*, 317-324.
- (33) Spek, A.L., Single crystal structure validation with the program PLATON, *J. Appl. Cryst.*, **2003**, *36*, 7-13; Linden, A., Adventures of a data validation editor, *Acta Cryst.*, **2002**, *A58* (Supplement), C58; Spek, A.L., Automated detection of poor or incorrect single crystal structures, *Acta Cryst.*, **2002**, *A58* (Supplement), C58.
- (34) <http://www.rcuk.ac.uk/access/index.asp>
- (35) <http://www.nih.gov/about/publicaccess/index.htm>
- (36) <http://www.reciprocalnet.org/>

---

(37) Lagoze, C.; Van de Sompel, H.; Nelson, M.; Warner, S., The Open Archives Initiative Protocol for Metadata Harvesting, Version 2.0. **2002**, <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>

(38) Duke, M.; Day, M.; Heery, R.; Carr, L.A.; Coles, S.J., Enhancing access to research data: the challenge of crystallography Proceedings of the 5th ACM/IEEE-CS Joint Conference on Digital Libraries, **2005**, 46 – 55, ISBN:1-58113-876-8