DIGITAL REPOSITORIES SUPPORTING ERESEARCH: EXPLORING THE ECRYSTALS FEDERATION MODEL. EBANK/R4L/SPECTRA JOINT CONSULTATION WORKSHOP, LONDON, 20TH OCTOBER 2006

A report by Dr. Wendy A. Warr

Wendy Warr & Associates

December 2006

Dr. Wendy A. Warr
Wendy Warr & Associates, 6 Berwick Court
Holmes Chapel, Cheshire CW4 7HZ, England
Tel/fax +44 (0)1477 533837
wendy@warr.com http://www.warr.com

CONTENTS

Digital Repositories Supporting eResearch: Exploring the eCrystals Federation Model.
eBank/R4L/SPECTRa Joint Consultation Workshop, London 20th October 2006

A transcript by Dr. Wendy A. Warr

Presentations

Introduction, Background and Context
Liz Lyon, Director of UKOLN, University of Bath, E.Lyon@ukoln.ac.uk

Liz Lyon is Director of UKOLN where she supports the development and implementation of the Information Environment, promoting synergies between digital libraries and eResearch. She has led the eBank UK project, and is Associate Director (Outreach) of the UK Digital Curation Centre (DCC) in which UKOLN is a partner.

Electronic publication has become the route of choice for dissemination and discovery of scientific works, but this remains simply a mechanism for the process, and the structure and content of an electronic article is largely the same as that of a paper version. Thus the release of scientific data into the public domain is constrained in exactly the same way as it was 10, 20 or even 30 years ago. The current rate that data may be generated and captured therefore far outweighs the rate of dissemination.

The eBank UK project, funded by the Joint Information Systems Committee (JISC) and progressed in two phases since September 2003, has investigated the Open Archive Initiative (OAI) approach as a solution to this problem, and the linking from primary data to other research outputs within the scholarly knowledge cycle. Building on the OAI concept, the project focused on chemical crystallography and constructed an institutional repository, eCrystals that makes available the raw, derived and results data from a crystallographic experiment. Following the creation of a completed crystal structure, data are uploaded into a data repository and additional metadata (chemical and bibliographic), to Dublin Core standards, are associated with the data set. This approach allows rapid release of crystal structure data into the public domain, but can also provide mechanisms for value-added services that allow discovery of the data for further studies and reuse, whilst ownership of the data is retained by the creator.

For a repository to be interoperable with other repositories, *via* an integrated research infrastructure, and to enable a harvesting process by third party services, the repository must publish its metadata according to a strictly controlled schema. eBank UK has developed a metadata application profile for the crystallographic data repository, which has been approved by the crystallographic governing body, the International Union of Crystallography (IUCr). All crystallographic data conventionally published in journal articles are collected by the Cambridge Crystallographic Data Centre (CCDC) and made available as the Cambridge Structural Database (CSD) and CCDC has agreed to harvest data from institutional data repositories for incorporation into the CSD. Journal publishers in the chemistry domain, such as the Royal Society of Chemistry (RSC), IUCr and Chemistry Central, have expressed interest in adopting the eBank UK model for the publication of primary scientific data so that the data may be cited and linked to a formal article.

Phase 2 of the eBank project is coming to a close and Phase 3 is to be launched. Phase 3 is an eight-month project that will progress the establishment of a global federation of data repositories for crystallography by performing a scoping study into the feasibility of constructing a network of data repositories: the eCrystals Federation. The federation approach builds on the work of the

eBank project and has links to three other projects in chemistry: Repository for the Laboratory (R4L); Submission, Preservation and Exposure of Chemistry Teaching and Research Data (SPECTRa) and Smart Tea (an electronic laboratory notebook study). The federation will contribute to the development of a digital repository infrastructure for research.

In Phase 3, partners will work together to harmonise the metadata application profiles from repositories operating on different platforms (ePrints, DSpace and Reciprocal Net); investigate aggregation issues arising from harvesting metadata from repositories in other countries; and scope the issues of the federation of institutional archives interoperating with an international subject archive (IUCr). The IUCr subject archive will primarily be concerned with data preservation, and provision of a facility whereby any researcher can openly deposit data, whilst CCDC harvesting will further populate the existing CSD, and the eBank aggregator service will address the issues of linking data sets with primary sources of publication.

Phase 3 will also explore data curation and preservation issues, advocacy, and sustainability within a federation. The project extends a successful working partnership that unites members of the digital library and eResearch communities. This collaboration now formally includes participation from digital preservation experts (DCC and Council for the Central Laboratory of the Research Councils (CCLRC)), publishers (IUCr, RSC and Chemistry Central) and the key professional and commercial bodies in the field (IUCr and CCDC). The project will continue to be led by UKOLN with core partners at the University of Southampton and DCC. Intute at the University of Manchester was a core partner in eBank but is a supporting partner in Phase 3.

One object of the October 20th workshop was to gain feedback from some of the supporting partner organisations: IUCr; CCDC; RSC; Chemistry Central; SPECTRa at the University of Cambridge; CCLRC; University of Sydney, Australia; and Reciprocal Net at the University of Indiana. The point of contact with each partner is to be identified. The eBank Web site will be updated with links to collaborators and partners, and a discussion list will be set up.

Institutional Data Repositories for Chemistry
Simon Coles, School of Chemistry, University of Southampton, s.j.coles@soton.ac.uk

Simon Coles is Manager of the UK National Crystallography Service (NCS) and has led the NCS involvement in the UK eScience programme from the CombeChem project. He is a founder member of the CrystalGrid Collaboratory and a co-investigator on the digital repository based projects eBank UK and R4L. Partners in the eBank project team were UKOLN; the Intelligence, Agents, Multimedia Research Group (formerly the Multimedia Research Group) in the Department of Electronics and Computer Science at the University of Southampton; and the School of Chemistry at the University of Southampton, which hosts the National Crystallography Service. The eBank UK project has produced a prototype demonstrator of a service (based on ePrints.org software) providing access to the detailed results of scientific experiments in crystallography.

Why is there a need for institutional data repositories for chemistry? In a news release dated 28th June 2005 (http://www.rcuk.ac.uk/news/20050628press.htm, and statement updated at http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf), the funding body Research Councils UK (RCUK) stated that the UK government takes the view that the data underpinning the published results of publicly funded research should be made available as widely and as rapidly as possible. So, JISC has raised the issue of institutional data repositories. As scientists, we should be worried about curation in the laboratory. Scientists need better ways to re-use data to vet experiments, correct errors, or make innovations, yet data from even recent experiments may be irretrievable or undecipherable. Unlike the data from “Big Science” (e.g., genomics and climate research), data from “Small Science” are heterogeneous, hugely numerous and increasing at a faster rate (Carlson, S. Lost in a Sea of Science Data. *The Chronicle of Higher Education* 23rd June 2006 http://chronicle.com/free/v52/i42/42a03501.htm).

Approximately 30 million chemical compounds have been made and some 1.5 million crystal structures have been measured. Of these, only 450,000 are held in the databases maintained by three or four data centres. The whole process of publishing such data is fraught by information loss. A PhD student will generate several spectra every day, many of them to check the progress of reactions etc., but only the spectrum of the final product is likely to be published. Even if it is published, it will be reduced to one hard copy figure with some of the peaks listed in the accompanying text. A great deal of information is lost in the publishing process.

In the approach used at the University of Southampton, data are separated from their interpretations but the two functions are linked. The interpretation and intellectual dissertation appear in a journal article or report, while the underlying data are held in an institutional data repository. In this way, *all* the data can be made available. The R4L project aims to capture and curate data at the point of generation in the laboratory. The project team wants to change the mindset of researchers, and capture the data, and provenance, as they are generated, not retrospectively.

The first procedure needed is an interface to the laboratory instruments. This involves collaborating with instrument manufacturers to develop protocols for data deposition and metadata; and devising a service to establish a reliable timestamp to provide a legally sound guarantee of priority. Next, management protocols and tools are designed to manage multiple, heterogeneous data sets in a repository. A laboratory repository can then be built and linked to external repositories by means of OAI. A tool to generate formal descriptions of the experimental process and compare data from different analyses allows scientific data reports, such as journal articles and preprints, to be prepared using OAI.

The reports can be stored in an institutional repository. The R4L project team has collaborated with the Association of Learned and Professional Society Publishers (ALPSP) and the eBank UK project to develop data citation and aggregation protocols. Data citation reporting is then possible using the institutional repository and OAI. Data dissemination and aggregation, through eBank, can be carried out using the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). Scientists can thus re-use data in a scientific instrument, completing the data capture and reuse cycle. The R4L researchers have built a prototype repository and are starting to populate it with experimental data and metadata. They are starting to assemble, automatically, records of experiments that can be downloaded. A reporting tool is being developed by R4L as a spin-off of the eBank project. It pulls together data sets from different analytical techniques to provide a report that is accepted by the publishers, can be deposited in an institutional repository and maintains the links back to the original data.

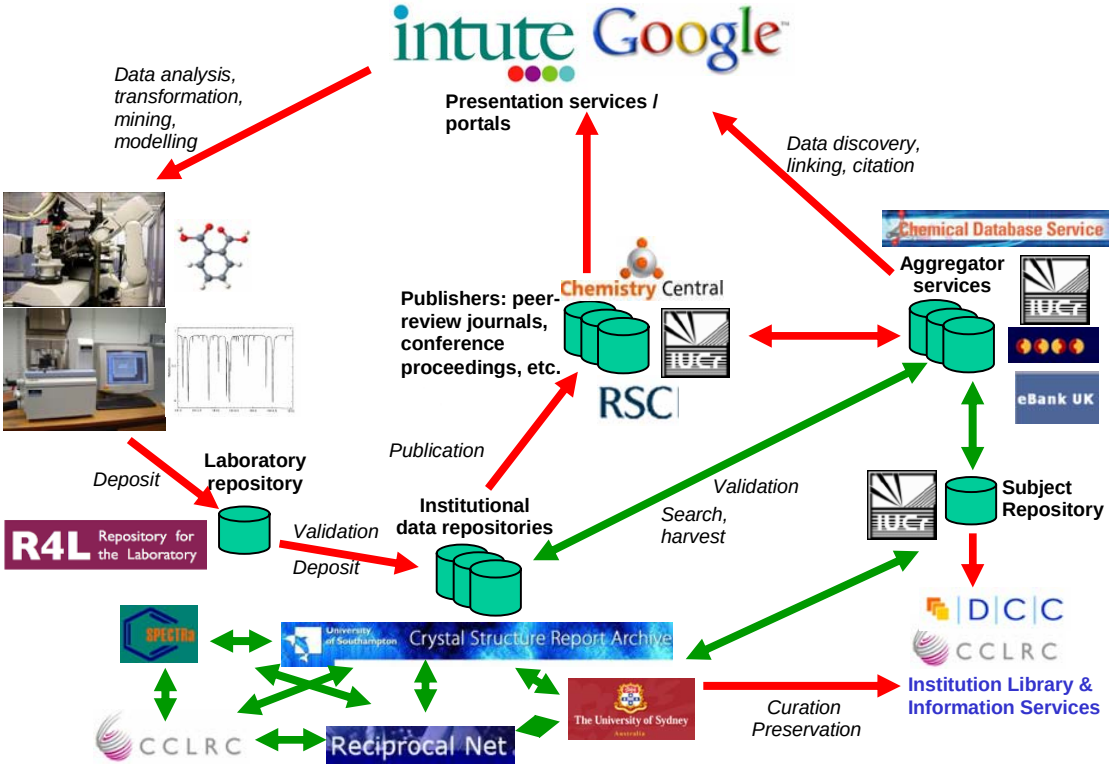
The eCrystals repository makes available the raw, derived and results data from a crystallographic experiment. Coles showed record 145 from the Crystal Structure Report Archive (http://ecrystals.chem.soton.ac.uk/145/). A rotatable 3D structure was displayed alongside data such as the molecular formula, the chemical name and the International Union of Pure and Applied Chemistry (IUPAC) International Chemical Identifier (InChI). Data collection parameters and refinement results were displayed. Many files for downloading were also listed. Coles emphasised the fact that all available data are being made available, including value-added data and audit trails. The repository is not yet directly interfaced to an instrument but a metadata schema has been developed and the project team is experimenting with Digital Object Identifiers (DOIs), keywords, classes and ontologies.

To make eCrystals data available to the public through OAI, metadata are published. Simple Dublin Core is used for the crystal structure, title (i.e., systematic IUPAC name), authors, affiliation and creation date. Additional *chemical* information (empirical formula, InChI, compound class and keywords) is published through Qualified Dublin Core. The metadata specify which "data sets" are present in an entry. The eCrystals entry shown as an example has DOI http://dx.doi.org/10.1594/ecrystals.chem.soton.ac.uk/145 but there is a monetary cost associated with the use of DOIs and there are technical issues to be overcome. Intellectual property rights

are defined at http://ecrystals.chem.soton.ac.uk/rights.html. Metadata schemas are listed at http://www.ukoln.ac.uk/projects/ebank-uk/schemas/. Quality control is applied to the archive to ensure that, as far as possible, the user gets a full and correct record. The Crystallographic Information File (CIF) is checked (a checkCIF feature is incorporated in a data manipulation toolbox), value-added data are handled, formats can be converted, and all associated metadata are stored.

It turned out that just one archive was not sufficient. Initially, eCrystals was a dissemination tool but the team now recognises the need for a repository behind a firewall, i.e., a "green" archive, held private. There are timing issues, and public/private considerations, in dissemination and the interface with journals. Roles have to be considered. All these issues were studied in Phase 2 of eBank.

The project is now moving into Phase 3, looking at institutional data repositories, and harvesting, aggregation and curation by data centres and third party services. The requirements are being captured for eBank UK Phase 3, the eCrystals Federation. A system has to be rolled out and all the stakeholders have to be consulted first. Coles presented this diagram of the eCrystals "global federation" model, with a caveat that the diagram is still a work in progress.



Phase 3 will explore the heterogeneous landscape of data repositories: different software platforms such as SPECTRa and Reciprocal Net; different administrative domains such as the University of Indiana (for Reciprocal Net) and the University of Sydney; different institutional structures (e.g., CCLRC); and different types of repository (subject repositories such as that of IUCr, and institutional repositories). New initiatives such as Object Reuse and Exchange (ORE) will also be considered. ORE is a new, two-year effort by the Open Archives Initiative, that began in October 2006. ORE will develop specifications that allow distributed repositories to exchange information about their constituent digital objects. The project is supported by the Andrew W. Mellon Foundation and co-coordinated by Herbert Van de Sompel and Carl Lagoze.

Repository entries will be harvested by established data centres, such as CCDC, and other aggregator services that link data to publications such as the eBank prototype. Preservation and curation by data centres and institutions e.g., DCC, and the Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval community (CASPAR), will also be considered. The gigabytes of data gathered about a structure from an instrument result in megabytes of data in eBank and IUCr reports, which are in turn related to kilobytes of data in a CCDC record.

Another factor is the relationship with "conventional" publication protocols and procedures, e.g., IUCr discipline-based publication, RSC domain-based publication and Chemistry Central Open Access publication. Last but not least, aggregation, linking and information provision by third party services will be considered. Functions include indexing (e.g., in Google), aggregating with other data sets such as those in the Chemical Database Service; aggregating and linking between data sets and articles, as in eBank UK, and integration into information portals such as Intute.

The Discovery Landscape in Crystallography

Monica Duke, Software Developer, UKOLN, University of Bath, m.duke@ukoln.ac.uk

Duke's talk described where digital libraries fit into the discovery landscape in crystallography, and potential problems for discovery. In the e-mail era before the appearance of the World Wide Web, there were a small number of communications, one-to-one, in a tightly managed, trusted, network. In the Web era we have agreed formats for exchanging crystallography data and the data are linked to journal articles. In the more independent, distributed infrastructure of the Semantic Web of the future, it will not just be a case of the final result being published in an article. The chemistry department at Southampton has reported on the publication@source model (http://eprints.soton.ac.uk/1633/). In this model, an e-print makes available all raw, derived and results data from a crystallographic experiment *via* a searchable and hierarchical system. At the top, searchable level these metadata include bibliographic and chemical identifier items which allow access to a secondary level of searchable crystallographic items which are directly linked to the associated archived data. Hence the results of a crystal structure determination may be disseminated so that anyone wishing to use the information may access the entire archive of data related to it and assess its validity and worth.

In the Semantic Web era, individuals will also put materials on Web sites and publish articles in Open Access journals. On top of the publication@source information and the published articles, will be services such as ChemRefer (http://www.chemrefer.com), DAREnet (a service that gives free access to academic research output in the Netherlands), and the über service" OAIster (a metadata harvester that provides a broad, generic retrieval resource for information about publicly available digital library resources).

Repositories and related services may be managed by individuals, institutions, or professional societies. Protocols, and procedures such as level of control vary, as do the formality, documentation, and comprehensiveness of policies. Coverage may be subject-related or may be national or international. Many of these services are designed for use by humans, rather than computers. There may therefore be incomplete support for automated information exchange and agents. Web interfaces and searching capabilities differ, and, with the increase in the number of services, users cannot search each repository individually. Furthermore, some new search facilities such as InChIs do not lend themselves easily to human manipulation in a Web interface and are better suited to automated processing.

In digital library infrastructures and technologies, service providers use harvesting based on OAI-PMH to gather data from multiple data providers. The OAI-PMH, currently at version 2.0, is an OAI protocol for metadata harvesting. It is a simple protocol based on HTTP, XML, XML schema and XML namespaces, for sharing metadata records between applications. It allows a harvester to ask a remote repository for some or all of its metadata records. Which data are supplied

depends on date-stamps, sets, and metadata formats. Harvesting is incremental and the data are partitioned into sets.

Metadata in the eBank UK project are defined according to the Simple Dublin Core standard. It is intended for resource discovery, compatible with OAI-PMH, and qualified to specify "vocabularies". Refinements aid interpretation of element value. An example is

<dc:subject xsi:type="ebankterms:CompoundClass">Organic</dc:subject>.

The metadata terms creator, rights, date, type, identifier and subject (including InChI and chemical formula), and others, are specified using XML schemas and documented using an application profile (http://www.rdn.ac.uk/oai/ebank/20060310/ebank_dc.xsd and http://www.ukoln.ac.uk/projects/ebank-uk/schemas/profile/).

OAI-PMH is only a partial solution; the eBank infrastructure needs to encompass other technical solutions. Other problems of OAI-PMH are immature experience of service provider models; identification of repositories of interest and the subset of the content in them; duplication of resources; and metadata quality. What makes good metadata and how can they be generated consistently? The definition of "good" depends on what you want to use the metadata for. Should metadata generation be mostly automated? *Can* it be mostly automated?

Duke produced a long list of questions to be answered:

How can these publications in different repositories (e.g., DAREnet, and chemical databases) and the data be joined up to offer useful services to users?
What is the role of OAI-PMH?
What other interfaces need to be considered?
Which communities use crystallography data?
How can the communities of users be defined and described?
What is a "useful" service?
Do users have overlapping information needs, and an interest in common subsets of sources?
How can information needs be identified and described?
What sorts of solutions are appropriate?
What are the interface design implications?
What discovery tools are already being used?
Can tools and services be adapted, or do we need new ones?
What is the role of publishers?

Some information sources of use to crystallographers are cross-discipline (e.g., OAIster and DAREnet). Others are discipline-specific: ChemRefer and Chemistry Central are concerned with texts and publications in chemistry in general; the Crystallography Open Database (COD) and Reciprocal Net are concerned specifically with crystallography data. Some sources are metadata-based, within the OAI-PMH infrastructure. All have a variety of search interfaces, from simple to advanced. Well established sources include the Cambridge Structural Database and the Protein Data Bank (PDB).

OAIster is an OAI-PMH aggregator. It is wide-ranging and inclusive, for any repository, and all content types. It reads metadata from 675 institutions. A keyword search (e.g., for "crystallography") can be limited by resource type (text, image, audio, data set etc.). OAIster finds five data sets in crystallography; the results give pointers to collections of data. Entering "crystallography" in the simple (not the advanced) search interface i.e., not limiting the search by the resource type "data set", yields more than 2000 records. The results are spread across several sources.

DAREnet gives worldwide access to Dutch academic research results. A simple search on "crystallography" yields 40 results. A general, advanced search (on author, year) can be carried

out. ChemRefer has a simple search interface for accessing full text chemical and pharmaceutical literature. Chemistry Central currently has no search feature but it is searchable through BioMed Central. The Crystallography Open Database promotes open data. It allows submission of CIF files; alternatively, condensed data can be deposited in a simple "REF" format. COD has about 43,000 searchable entries. Reciprocal Net is a distributed crystallography network for researchers, students and the general public. It has a search engine with a crystallography-specific interface.

Semantic Interoperability in a Federated Crystallography Information Service
Traugott Koch, Research Officer, UKOLN, University of Bath, t.koch@ukoln.ac.uk

There is a need to enhance interoperability amongst data, images, text and metadata (both open and proprietary, free and fee-based) in data repositories (institutional, disciplinary, national, and international), publication repositories, data and publications on the Internet, aggregators, databases and services. Koch concentrated on semantic interoperability as opposed to syntactic interoperability. Syntactic interoperability is about applying common formats and protocols for data transfer and merging (e.g. CIF, XML, OAI, Z39.50, SRW). Semantic interoperability is about shared meaning of the content.

Approaches to enhance semantic interoperability include agreed common standards at all sites; conversion and normalisation; mappings between site-specific solutions; and enhancement of metadata with vocabularies, schemes, mapping, and names. When and where these approaches are applied depends on the selected architecture. Three areas need to be considered: data structures such as metadata profiles; categorical data such as topics and classification; and factual data, e.g., names, formulae, and other named entities. For full semantic interoperability, all three areas need to be addressed. Koch discussed each in turn.

A multiplicity of metadata profiles is in use, some of them documented and some not. eBank has implemented the following metadata engineering:

- The eBank data model (http://homes.ukoln.ac.uk/~tk213/pres/ebank-model.ppt)
- The eBank metadata application profile http://www.ukoln.ac.uk/projects/ebank-uk/schemas/profile/ and http://homes.ukoln.ac.uk/~tk213/pres/ebank-AP.html
- Namespaces, i.e., metadata terms (http://www.ukoln.ac.uk/projects/ebank-uk/schemas/terms/ and http://homes.ukoln.ac.uk/~tk213/pres/eBank-terms.html)
- The XML Schema specification for eBank terms **[This link not yet live on 18th December, 2006]**
- eBank metadata output in Metadata Encoding and Transmission Standard (METS) format under the OAI protocol; and oai_dc formats for the OAI harvesting protocol (http://ebank.eprints.org/perl/oai2/?verb=ListRecords&metadataPrefix=ebank_mets and http://ebank.eprints.org/perl/oai2/?verb=ListRecords&metadataPrefix=oai_dc)
- Data dictionaries by applying the Crystallographic Information Framework (http://www.iucr.org/iucr-top/cif/).

Koch showed extracts from the metadata application profile, which determines how the elements will be combined and used, and from the OAI harvesting files. The example was the use of a chemical formula in eBank.

Potential actions for preparing a crystallography federated data repository in eBank Phase 3 are agreeing on a metadata solution including value encoding for a future service; developing a common application profile for crystallography data; defining different degrees of interoperability and adherence to the common model, including a minimum level; and taking steps towards harmonisation with the profiles of related publication servers.

Categorical data can take the form of subject headings, keywords, classification, etc. This field is underdeveloped in crystallography. eBank currently uses controlled keywords (http://ecrystals.chem.soton.ac.uk/key_A.html) adapted from the IUCr online World Directory of Crystallographers (http://www.iucr.org/iucr-top/wdc/index.html. "Compound class" is a classification eBank uses. Since it consists of only four or five categories, it is useful only for filtering. There are too many hits for each class when browsing or searching one class without additional search argument, even in a small database such as the eBank repository. eBank does not currently use the keywords used in the IUCr journals *Acta Cryst. A*, *Acta Cryst. B* and *J. Appl. Cryst.*; or the keywords and classification used in the 25 abstract and indexing databases (http://journals.iucr.org/services/abstracting.html) which cover the IUCr journals in Crystallography Journals Online (http://journals.iucr.org/). All these keywords could be used for creating a future eBank or crystallographic controlled vocabulary. Are there other sources of vocabulary in the area of crystallography?

Potential actions for eBank Phase 3 are clarifying the data model question (whether the keywords characterise elements of the data or the context of the problem or use); developing and maintaining a common controlled keyword system and classification for crystallography, with coverage of data-related topics; and developing and maintaining mapping to related discipline and generic classifications.

Koch next discussed the third area: factual data, and named entity standards and authorities. Author and institutional names are currently used in uncontrolled and unvalidated form. Names in the World Directory of Crystallographers and the IUCr ID (the username used to access the directory), and in library- and university-related name authority projects, and national projects are other potential sources for creating a name authority list for eBank, now or in Phase 3. "Names" for the objects of study and their components (crystal structures, chemical compounds etc.) include IUPAC chemical name (standardised in the so-called "colour books" http://www.iupac.org), InChI, and chemical formula (based on CIF). What names are used by other organisations in the future federation?

Again, Koch listed potential actions for eBank Phase 3. The team members could co-operate in building and improving name authority databases, contribute to further standardisation (e.g., InChI, and the IUPAC Compendium of Chemical Terminology known as the Gold Book http://goldbook.iupac.org/) and use name authorities for verification and metadata enhancement. They could build authorities into metadata creation tools (e.g., the repository submission toolset in DSpace or ePrints), for example, *via* Web services. They could also experiment with mashups of crystallography data to investigate what should be the candidate reference schemes and to determine the consequences for repositories.

Koch listed a number of unanswered questions regarding standardisation. To what degree can existing conventions be seen as standard? Does the convention or standard lead consistently to the same "name"? Should eBank use proprietary systems such as Chemical Abstracts Service (CAS) Registry Numbers (CASRN)? There are also some general problems with common and standardised solutions, e.g., standardisation processes, adoption, validation, and maintenance.

Finally, Koch discussed some application issues. In metadata creation and subject assignment, eBank could use the same tools as the ones supporting discovery (e.g., semi-automated indexing), or similar tools. "Cataloguing" rules in accordance with the data model and the application profile might be "harmonised". The benefits of text and data markup (e.g., Chemical Markup Language, CML) in combination with keyword indexing and full text searching could be explored. Experiments with participatory indexing (social tagging, folksonomies) could be carried out by research groups in specified research areas. The need for formal ontologies and logical reasoning over the data and the literature in crystallography could be investigated.

Discovery (i.e., searching, browsing and linking) raises other application issues. Searching in components and substructures can be done using strings or with graphical search support (e.g.,

JChem and Marvin, http://www.chemaxon.com/). Searching of (and filtering with) key characteristics of crystal structures (as in Reciprocal Net and the Crystallography Open Database), searching inside the data files and the CIF format, and searching in information services with much broader topical coverage (university-wide repositories, national data archives, OAIster, Google) all need considering. The team could also investigate the potential benefits of text and data mining for indexing and searching support and take steps towards knowledge extraction, hypothesis creation and larger-scale computational processing.

Koch's presentation is available on the Web at http://homes.ukoln.ac.uk/~tk213/pres/eBank-ws200610.html. His eBank terminology report, "Terminology and subject access issues regarding eBank UK" is at http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/termino-public.html.

Research Data and E-learning: Findings from the Evaluation of eBank
Gráinne Conole, Institute of Educational Technology, Open University, g.c.conole@open.ac.uk

Gráinne Conole did a PhD in crystallography and now does research in e-learning. Her final report on the evaluation of eBank is currently in preparation. It focuses on the lessons emerging from eBank, a comparison with related initiatives, and the implications for research, teaching and policy. Methods used in the evaluation were documentary analysis, interviews with key members of the project team, and observational analysis of and interviews with MChem students at the University of Southampton. The project findings cover project aspirations, collaboration and inter-disciplinarity, links with related projects, key success factors and outcomes, dissemination mechanisms, barriers and enablers, conceptual models and pedagogical issues, the student experience, and future directions and recommendations. In the talk at the workshop, Conole concentrated on links with related projects, key success factors and outcomes, barriers and enablers, conceptual models and pedagogical issues, and the student experience.

eBank fits into a complex set of projects. One group is Smart Tea, CombeChem, DAREnet, ARROW (an Australian repository project), Reciprocal Net, COD, CLADDIER (similar to eBank but for atmospheric data), R4L, Source-to-Output Repositories (StORe), SPECTRa, and Geospatial Repository for Academic Deposit and Extraction (GRADE). eBank also has wider e-science and e-social science aspects. It links to repositories such as ePrints and digital libraries, and also to e-pedagogy projects such as eMalaria, and the learner experiences of e-learning projects LXP and LEX.

Chemists, information scientists and computer scientists in a team share the same problem space but look at it from different perspectives. Interdisciplinarity is becoming more and more important. Conole exemplified this with two quotations:

"*And still I think there was an ambition to more systematically cooperate...e-Science, digital library and IT people…quite a hard thing to do…there are so many different interests and conceptions...the terminology is so different*."

"*There was the concept of thinking about data as a collection – collections of images, collections of books, but thinking of data as collections that you could describe and make available…*"

She gave two more quotations relating to the ingredients for success:

"*A meeting of cultures, which is quite an output in itself. Good to work at interface sometimes*."

"*It was about making connections and seeing where activities that are happening in one area can be migrated, transformed and transferred into another area*."

The vision and the demonstrator are linked. Success factors include shared visions, "triggers" (the people in this community knew each other already), track record, culture (crystallography is

an excellent area in which to work), stakeholders (crystallographers are able to tap into publishers and others), and dissemination (evangelising about the vision).

As for barriers, a comment from a respondent in the study was:

"*…main barrier is a socio-political one…it's a matter of a change of cultures…some people are embracing that, as I have said, others are a bit stand off-ish…you know, they like the mystique of publishing and that sort of thing*."

There are many other barriers. The concept of ownership is one: users are sensitive about what they see as "my data". Research practice is different in different areas and information is lost when someone leaves an organisation. The level of information and communication technology (ICT) skills can be a problem; crystallography is a good place to start because ICT skills are high in that discipline. The institutional infrastructure may or may not support institutional repositories. If repositories *are* supported, will they be supported in 10 years time? Southampton has a long history of supporting institutional repositories. Publishers' attitudes can be a problem: will the publishers make material available? Technical and funding issues and competing agendas are further problems. Intellectual property rights are a big issue in the move towards more openness, both open source and open content. The creative commons licence has been proposed. There is an open content initiative at the Open University. Three aspirations of openness are:

- To make data available through open access, so that they could be disseminated more quickly
- To link data to derived references and enable demonstration of provenance
- To see research data available and applied in the learning context as a means of completing the scholarly knowledge cycle.

eBank contributes to the scholarly knowledge life cycle by linking teaching, data, and research with publication. Conole is interested in taking this forward into nascent pedagogical models. There is much interest in "research-led teaching" nowadays. This is a very new field. The Chemical informatics module (6016) at the University of Southampton consists of a Blackboard site plus eBank plus eMalaria. Pedagogical benefits of this sort of module could include improved ICT skills, wider use of ICT, and other desirable results. Comments from students include the following (unedited, apart from apostrophes):

"*It's quite an interesting course, it's quite different to a lot of courses I've done… basically it's comprised of …you had set lectures, then you had workshops and you also had this kind of an assignment which was very much kind of do it yourself. It's like what was a project … at the time when I first got given it I didn't think it would take up as much time as it did. I mean really it did take up a lot of time*."

"*There were several parts to the course. We started off with how to get 2D and 3D representations of molecules onto a computer using a one-dimensional format, a SMILES string …so just ways of like getting data into a format so that it can be easily shared between different computers or different people without having to change lots of things*."

"*Another quite nice section of the course, involved databases … searching databases and getting more use out of databases and how the best way to go about this and also how to put information into a database so if you come up with say a crystal structure...how to get that into a format that the database will accept so that it's easily accessible by lots of other people*."

"*I think basically for me it's being clarifying and really actually now understanding things I've been using for a while. Something like linear regression...I used without really understanding, whereas now … I understand now it's not complete magic … yeh hands-on experience …now understand a bit about it, wouldn't say I understand it completely, but given me a better understanding*."

"*Before the course I hadn't really considered how the computer actually does it but … interesting to see how that works and then there is a part of the course where…they taught you…how to interpret data and build models…and that's probably quite a useful part of that project use the model building…it's all very well people telling you this is a peptide but until you actually use it, you can't really visualise it*."

"*Well basically I've done nothing like it before, so it's the first time I've sort of delved into computing or computational chemistry…quite nice, quite enjoyed starting off with just like a string of data and pop it into say a database, just a flat string of numbers basically and then come out with a crystal structure, which is exactly what it should represent which is quite cool*."

"*I'm connected to broadband, constantly online, [that's at home?] yeh. I mean it helps using the Internet, researching, finding stuff out…and I think you can, there was a stage when you could write up your lab reports by hand but now it's basically presumed that people do it on the computer, so all assignments, all lab reports, everything is written up [on the computer]…which is good for me because I am pretty bad at spelling*."

"*But also personally I like to make a good presentation where I've got formulas and have a test…it's quite sad...as a test…so I have one slide saying Hess' law and the next slide...and I have to write it down [you use it to do your own tests?] yes exactly it's the only way I learn*."

"*Being able to communicate with the lecturer…I e-mailed [the lecturer] on Saturday night at eight o'clock and he replied at midnight…. I don't know what that says about him or me, but…certainly e-mail…easy to get in touch with the rest of the class…and him*."

"*Just started using that actually [MSN chat] [with friends?] yes yes, I got told to use it…'I've had enough of phoning you up it's costing me a bomb get MSN it's free'*.

"*I use the Internet a lot…to do…research, something I don't understand I might have a look on the Internet and see the different explanations to help.*"

"*…Strathclyde University I think, one of their guys had done a basic summary of like regression and … help you understand the basic principles behind…not just for this course but for the course in general...In Google, typed in…linear regression.*"

Students were enthusiastic about module 6016 and saw it as different from other courses. They accessed real data and were able to *understand* techniques such as linear regression. They appreciated the hands-on experience and the sense of "using real stuff". This was a case of understanding by doing, and authentic learning. These students use technology all the time: notice the comments about broadband and everything being written up on the computer. The spelling issue came up quite a lot. The Internet is the student's first part of call, especially Google and Wikipedia. In this case, the students are using eBank and other databases. Conole is working on another, broader project (LXP) looking at the use of technology. In both module 6016 and LXP the students are taking technology on board. Students are personalising the technology: note the student who e-mailed a lecturer and was impressed by the response. Use of other Web sites was also helpful: note the mention of linear regression at Strathclyde.

Research in e-learning is usually focused on technological or pedagogical issues and so far students have been largely overlooked, so LXP turned its attention to students who are using technologies to support their learning activities. In LXP, an online survey (http://www.geodata.soton.ac.uk/eLRC/learner_survey), audio logs, and interviews were used to study uses of technologies, effective e-learning strategies, subject discipline differences, and student experiences.

Students are using technology in lots of different ways, including social networking and so-called "Web 2.0" sites which harness the masses and Grid technologies. Students use mobile

technologies, Google, e-journals, podcasts, blogs, and wikis in an integrated, multi-faceted fashion. The UK research funding bodies have recognised this. The Higher Education Funding Council for England (HEFCE) Centres for Excellence in Teaching and Learning (CETL) initiative has two main aims: to reward excellent teaching practice, and to invest further in that practice so that CETL funding delivers substantial benefits to students, teachers and institutions. The National Centre for e-Social Science (NCeSS) is funded by the Economic and Social Research Council (ESRC) to investigate how innovative and powerful computer-based infrastructure and tools developed over the past five years under the UK eScience programme can benefit the social science research community.

If you talk to students, you will hear eBay, Wikipedia and YouTube mentioned again and again but they are used differently depending on individual needs. Important trends are shifts from information to communication, from individual to social, and from passive to interactive. Even listening to things is no longer a passive activity. All lecturers give out PowerPoint hand-outs nowadays and they annotate them.

ICT has become pervasive and integrated. The tools are used extensively for everything and are personalised, i.e., adapted to personal needs. The networked peer community uses social software. Content is not fixed but is created interactively. New skills are needed and they are arising. Students are using technologies to support all aspects of their lives not just learning. Therefore the skills and experiences they gain from the use of tools in these different contexts are being transferred into how they use them for learning, e.g., using MSN chat (http://groups.msn.com/) for social reasons and then for learning, using online booking sites, using community software such as delicious (http://del.icio.us/), YouTube etc. The concept of time has changed: this is the “now” culture. In turn, working patterns, and ways of thinking and doing things are changing.

The SPECTRa Update: a Wider Chemistry Picture. A Digital Repository for the Chemical Community

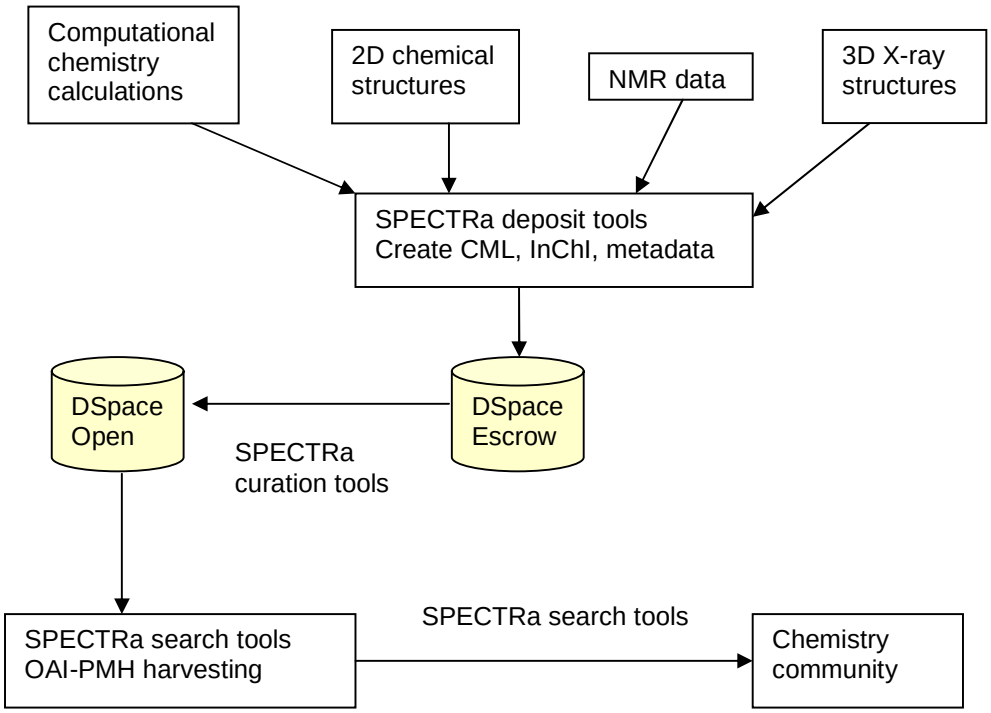
Alan Tonge and Jim Downing, University of Cambridge, apt24@cam.ac.uk, ojd20@cam.ac.uk

Alan Tonge is the project manager and Jim Downing is a software engineer at the University of Cambridge, in the Submission, Preservation and Exposure of Chemistry Teaching and Research Data (SPECTRa) project. This is an 18-month project between the University of Cambridge and Imperial College London to develop customised tools to deposit chemistry data in digital repositories. Libraries and chemistry departments are involved. SPECTRa is part of the JISC digital repositories programme and is closely integrated with eBank and eCrystals at Bath and Southampton. The object is to understand the needs of chemists and to provide tools for depositing data in a repository (in this case, DSpace).

User requirements in a number of different disciplines (synthetic organic chemistry, departmental crystallography services, and computational chemistry) are being determined by interview and by paper and electronic surveys. Crystallography is an ideal area in which to start because crystallographers understand added value and aggregation. The survey at Imperial is complete and the survey at Cambridge is now in progress. The second stage of the project will study specific data usage.

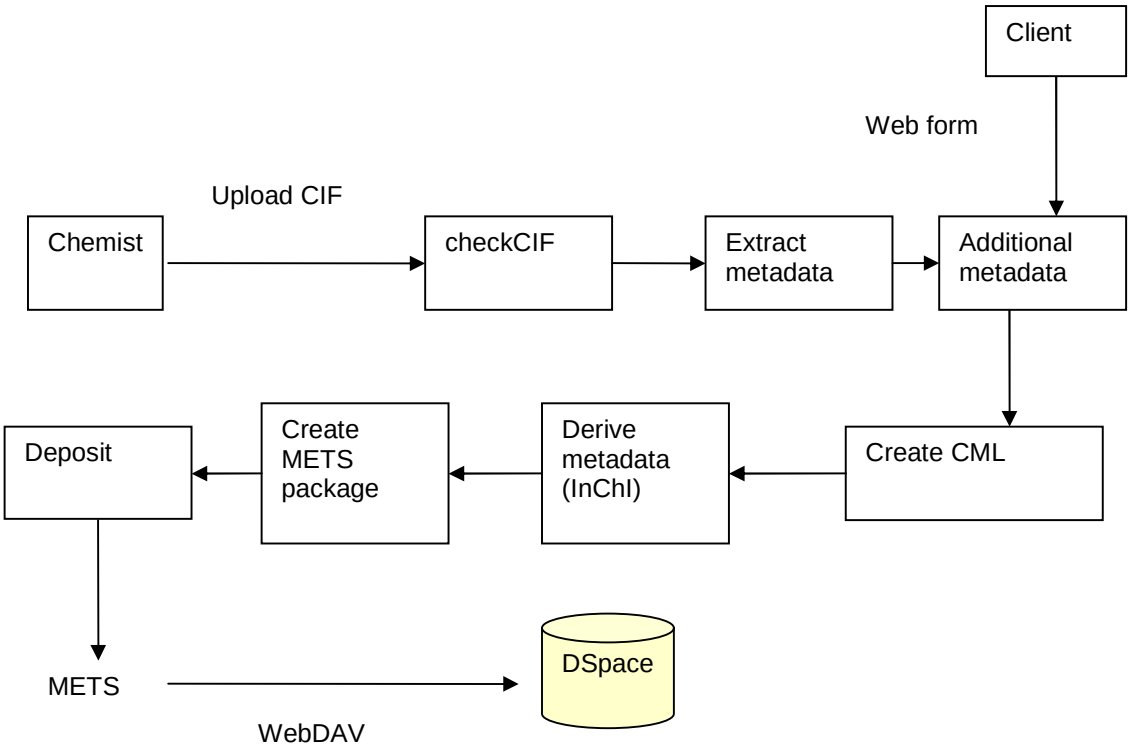
Science depends upon data. Experimental chemistry data are resources and assets but most of the data get lost or become unreadable. For example, proprietary formats for NMR, IR, and UV spectra have a five-year shelf life; supplementary data are often submitted to journals as PDF files that are not machine-readable; and 90% of CIF X-ray files remain unpublished. Let us assume that John Davies (a departmental crystallographer at Cambridge) has 3000 unpublished structures. At a cost of £300 per structure, this amounts to one million pounds worth of lost data. Most of the problems here are social, not technical.

Tonge and Downing outlined a solution. Selected data from chemistry workflows are captured in open format (say, a JCAMP-DX file, a Molfile, or a CIF); context-specific metadata and persistent identifiers are added; and everything is then deposited in a digital repository, for public release on the Internet. A new feature is *controlled* release, to meet the commercial requirements of funders, and the needs of chemists who have a phobia about open access. In controlled release, a period of escrow can be defined by the depositor. Once all the information is on the Internet, it can be found by search engines and OAI-PMH metadata harvesting. Diagrammatically, the architecture is represented as follows.



The institutional repository is not a domain focus: chemists deal with chemistry in their own repositories. The SPECTRa institutional repository is a “vanilla” one. It could be a cross-institutional repository. The RSC, or another organisation, might run the escrow service.

The SPECTRa crystallography submission tool works as follows.



There are so many standards: which should SPECTRa use as a packaging standard; in the deposit Application Programming Interface (API); for downloading a package or file; for the metadata harvesting API; for data in the OAI feed etc.? Data have to be moved around and ported: this is not a trivial exercise. OAI-PMH looks sufficiently well established but religious wars about identifiers could occur. The METS file manifest (list of contents) features were an obvious choice since it is the simplest API if DSpace is used. In WebDAV, identifying handles are resolved, packages are downloaded over HTTP, and Uniform Resource Locators (URLs) are resolved. The simplest tools have been chosen; tools that are not specific to crystallography or to DSpace.

For the escrow repository, there is a potential need to curate the embargo of content in a system other than the submission system, so the embargo details have to be written into the metadata. This has implications for a much wider field than that of crystallography alone. The legal niceties of the embargo licence, the post-embargo licence, and the time of duration of the embargo need to be established. Should the embargo be lifted manually or automatically? Perhaps the embargo should last for a maximum of three years before the data owner is consulted again.

Standards are like sausages: if you like them, do not watch them being made. Downing listed three standards which he thinks will be important in the field of federated repositories. The SPECTRa project should engage in the adoption of the JISC ePrints application profile, the JISC Deposit API, and OAI's Object Reuse and Exchange (ORE).

The Publisher's View of Crystallography Data

Richard Kidd, Manager, Editorial Production Systems, Royal Society of Chemistry, kiddr@rsc.org

RSC has mounted 6000 CIFs online since 1998. Most of RSC's supplementary information is crystallography files. These are freely available and are deposited with CCDC. From 2007, CIFs

will be checked programmatically with checkCIF as well as subjected to human review. More data will be moved from paper to CIF.

RSC has adopted Southampton's approach of separating the data and the interpretation but will keep CIFs with the paper in which the structure is published in a service hosted by RSC through which CIFs will be made freely available. OAIster might be used. From 2007 links to external data sources will be added. These will include links to additional relevant data stores, and data analyses such as properties, CML, and visualisation.

This integration demands longevity, which, in turn, requires permanence, accessibility, and use of DOIs. The tools for metadata availability and exchange are probably ready now but functionality and compatibility will need constant maintenance and updates, for example to allow for new browsers.

Publisher Perspective Two
Bryan Vickery, Deputy Publisher, BioMed Central, with responsibility for Chemistry Central, Bryan.Vickery@biomedcentral.com

Chemistry Central is a new initiative from BioMed Central, the open access publisher. Authors submitting to Chemistry Central journals will be encouraged to supply additional materials, such as crystallography data, spectra and visualisations, which the journal will deposit on their behalf in suitable open repositories. Each journal article will carry links to these data services.

BioMed Central fully supports the OAI Metadata Harvesting Protocol. Metadata for all the articles published by BioMed Central (including Chemistry Central) are available *via* the OAI interface. Additionally, thanks to BioMed Central's Open Access policy (http://www.biomedcentral.com/info/about/charter), repositories may also use the OAI interface to obtain the full text XML of any open access research article published by BioMed Central.

Publisher Perspective Three
Peter Strickland, Managing Editor, International Union of Crystallography, ps@iucr.org

Strickland has an interest in eBank from two viewpoints: that of IUCr and that of a publisher. IUCr is one of the scientific unions of the International Council for Science (ICSU). ICSU is a non-governmental organisation representing a global membership that includes both national scientific bodies (111 members) and international scientific unions (29 members). IUCr publishes eight primary research journals, overseen by its Commission on Journals (http://www.iucr.org/iucr-top/iucr/cj.html). It fosters cooperation between public curated databases (CCDC, ICSD, PDB, CRYSTMET, ICDD, etc., http://www.iucr.org/cww-top/data.index.html) through the Committee on Crystallographic Databases (http://www.iucr.org/iucr-top/iucr/database.html). IUCr promotes data exchange standards (e.g., CIF, mmCIF, and CBF/imgCIF) through its Committee on the Maintenance of the CIF Standard (COMCIFS). The IUCr is represented on the International Council for Scientific and Technical Information (ICSTI) and the ICSU Committee on Data for Science and Technology (CODATA).

The journals of the International Union of Crystallography are produced by the IUCr in Chester and published by Blackwell Munksgaard. The print editions of the journals are distributed by Blackwell Publishing, while electronic editions of all IUCr journals are available *via* Crystallography Journals Online (http://journals.iucr.org). The IUCr has published journals since 1948. Eight titles are currently published; the details for 2005 were as follows:

- *Acta Crystallographica Section A*, 6 issues a year, 700 pages
- *Acta Crystallographica Section B*, 6 issues a year, 1000 pages
- *Acta Crystallographica Section C*, 12 issues a year, 1500 pages
- *Acta Crystallographica Section D*, 12 issues a year, 1800 pages
- *Acta Crystallographica Section E*, 12 issues a year, 8000 pages

- *Acta Crystallographica Section F*, 12 issues a year, 1200 pages
- *Journal of Applied Crystallography*, 6 issues a year, 1100 pages
- *Journal of Synchrotron Radiation*, 6 issues a year, 600 pages

There are 16 editorial and administrative staff (16 full-time equivalents) and four research and development staff. Online services include Crystallography Journals Online (70,000 articles, 250,000 pages); World Directory of Crystallographers; checkCIF; and International Tables Online.

Crystal structure reports are a good example of primary research literature comprising detailed discussion of the quantitative results of well-defined experiments. The packing of atoms or molecules in the solid state within regular crystal lattices can be probed by experimental techniques such as X-ray diffraction. From the scattering data gathered in such an experiment, one may deduce much information about the nature and molecular structure of the components of the crystal, such as 3D positional coordinates, atomic motions, molecular geometry, chemical bonding, and crystal packing. Among the journals of the IUCr, two titles contain almost exclusively such structure reports (*Acta Crystallographica Section C: Crystal Structure Communications* and *Acta Crystallographica Section E: Structure Reports Online*). Such reports also form a routine component of longer research articles describing the chemical or physical properties related to structure, and so may appear in any crystallographic journal. Within the IUCr stable, such articles appear in *Acta Crystallographica Sections B: Structural Science; D: Biological Crystallography;* and *F: Structural Biology and Crystallization Communications*. In all these journals, scientific discussion and the presentation of the associated data are closely integrated.

The IUCr data archive consists of approximately 25,000 primary and 25,000 derived data sets. Strickland tabulated the nature of the data:

	Raw data (image plate, film)	Primary data (structure factors)	Derived data (structural model)
1948-1970s	none	print	print
1970s-1991	none	microfilm	print
1991-1995	none	microfilm	CIF
1995-present	none	CIF	CIF
Future	Archive	CIF	CIF

The IUCr is pleased that NCS at Southampton is archiving the raw data as well as CIFs and structure factors.

Strickland illustrated the importance of integrating journal and data publication. Although the editors and referees invest much effort in the peer review process, the availability of a full set of accompanying data provides added value for the reader of the article. For example, for any article published in *Acta Crystallographica Section E: Structure Reports Online*, the reader may assess fully the scientific argument by: (1) reading the text of the article; (2) accessing the full CIF (which will include unpublished data such as the three-dimensional atomic coordinates, and a complete listing of bond lengths and angles); (3) reviewing key indicators and the validation report (e.g., the author's response to a significant problem identified in the validation review); (4) retrieving the primary experimental data to allow an independent re-determination of the structure; or (5) visualising and manipulating the data in a crystallographic application of choice. Strickland showed a display from Structure Reports Online and a related display from a program, Mercury, distributed by the Cambridge Crystallographic Data Centre, allowing full visualisation of the six-dimensional model (displacement ellipsoids may be visualised as well as the three-dimensional positional coordinates), generation of the crystal lattice, exploration of molecular geometry and intermolecular bonding, and so on.

All data sets are checked with checkCIF so that the quality of the archive is assured. All data sets have DOIs; this is important since a DOI has permanence. Links are provided to structural data in Protein Data Bank entries, Nucleic Acid Database entries, and Cambridge Structural Database summaries. In future, links to other structural databases and federated data repositories will be added. Data are automatically deposited with the main crystallographic databases.

Focusing on the structures that sit on the edge of the publication spectrum, Strickland listed some of the possible outcomes. A number of chemistry journals will accept structural data sets as supplementary or supporting documents, and make these available from their Web sites, but others do not. Supplementary data files, where they are held, may or may not be fully compliant CIF files. (Fully compliant CIFs are important for information interchange and archiving.) Such structural data sets from journals that accept them are harvested by the curated databases (CCDC etc.) but substantial effort may be needed to retrieve and ingest the data. Authors may transfer their data sets voluntarily to the curated databases, but as with anything "voluntary", the coverage is patchy.

Increasingly, universities are setting up institutional repositories and encouraging their faculty members to deposit publications and (sometimes) research data. In principle, these repositories are open to harvesting if standard protocols such as OAI-PMH are used. There are domain-specific repositories where authors may voluntarily deposit their data; in crystallography, the Crystallography Open Database is available for any author, or, indeed, creator of unpublished structural data sets. In this discipline, there are also well-resourced initiatives to establish "data publication at source" through initiatives such as eBank and the Reciprocal Net. Valuable research carried out in the commercial world is unlikely to be published in research journals unless the overheads of creating such publications are very low. Structure factors are even less likely to be deposited with non-IUCr journals. If no option for data deposit is available, there is a real risk of total loss of valuable research data. A survey has shown clearly how much information can be lost when someone leaves an organisation or retires.

The aspects of eBank that are particularly important for long-term success are as follows. The initiative has some prospect of longevity especially when federated; it could offer security of operation through long-term funding arrangements. The eBank federation would use common protocols (e.g., CIF for the domain-specific data, OAI-PMH and METS for metadata dissemination and description, HTTP for content delivery, and DOI and OpenURL for identification and retrieval). eBank addresses domain-specific concerns. It is large enough (in terms of the federated entities) to discuss special arrangements for archiving (including discussions with publishers). Federation would foster resiliency, interoperability and common information management practices. eBank aims to be comprehensive within the user base and does not rely on voluntary action. It will facilitate transfer of data to curated databases and journals.

Strickland commended eBank for using standard data formats (CIF); for using OAI-PMH, DOI, OpenURL, and standard metadata; for providing links to all data and links to the related publication; and for handling issues such as rights and quality (the latter using checkCIF). In the short term, the IUCr can help by continuing to consult on metadata specification, and by advocacy through the Committee on Crystallographic Databases and CODATA. In the longer term, the IUCr could provide a Web index to data "publishers" such as eBank, perform validation analysis (checkCIF etc.), offer a search engine, and mirror and archive content.

ChemRefer. An Introduction

William Griffiths, CEO, ChemRefer, info@chemrefer.com

ChemRefer (http://www.chemrefer.com/) is a search engine for free, full text chemical and pharmaceutical literature. It supports Open Access articles in journals that are partially or fully on Open Access, i.e., almost all journals. Its simple text search interface is easy to use. The search engine supports technical terms and searches the *entire* article. It is indexed by Thomson's Web of Knowledge, it has authoritative links, and it has attracted interest from Open Access publishers

such as Arkivoc and Chemistry Central. The search engine spider carries out full text indexing but not full text copying. The text stays at the original URL, so accreditation is preserved. Most full text articles are stored as PDF files, which is ideal for publishers but causes a problem for search engines. The metadata are terrible and file sizes are huge. ChemRefer highlights search terms and synonyms within the resulting text, even in the PDF version, and it recognises, for example, that *Org. Lett.* is the same as *Organic Letters*.

Contributors who submit articles to ChemRefer are entitled to place their details on the ChemRefer contributors page, allowing them to promote themselves or their research through ChemRefer, free of charge. This facility is open to both commercial and academic researchers. Eventually, a network of researchers will be built up which could help instigate future scientific collaborations. Contributors (publishers, webmasters, authors or even interested Web surfers) can be anonymous if they like. ChemRefer tries to facilitate easier access through a toolbar, a newsletter and search on other sites. The toolbar allows a search box to appear in the user's own browser. (There is also a version for the Firefox browser.) The newsletter is a monthly e-mail linking to interesting articles indexed by ChemRefer, and examining other online chemical information resources. (A sample is available at http://www.chemrefer.com/ChemReferNewsletterAugust2006.html.) "Search on other sites" allows another Web site to have a ChemRefer search box on its site e.g., http://www.chemspy.com.

ChemRefer does not use default metadata, rather it customises and uses the metadata for title, author, citation, copyright, etc. The metadata are generated according to a submission form, one article at a time. Some publishers have shown an interest in "mass submission" of articles. Customised metadata are useful because they help in interpretation of search results and spider-fed chemical terms, and there is no need to alter the PDF. Search is a separate science from metadata. Metadata cope with chemical words, boost title keywords, for example, and optimise other keywords. Sources of metadata are the PDF of the article, publisher Web sites and data fed by publishers.

For each search result, search engines normally provide a few lines of metadata e.g., title and author, but they could provide many more. Why should they not supply 10 or 20 lines with much more comprehensive information? This is because searchers might want to look through potentially hundreds of results and if each search result were 20 lines long, this would take an impossibly long time. So, search engines must find a balance between getting the necessary information across (such as title and author names) to allow interpretation of search results, without overloading the searcher with so many data that a person cannot read through a set of results quickly and effectively.

Griffiths speculated on the possibility of manually adding extra chemical data, such as InChIs, to search results. It is not a service that ChemRefer currently plans to offer. The titles of articles in crystallography are different from those in other fields of chemistry. ChemRefer can handle crystallographic terminology.

In future ChemRefer wants to be involved in collaborative projects and to work with major publishers. The company is also looking for funding. [Postscript: since the workshop, ChemRefer has worked with IUCr to include IUCr Open Access articles in ChemRefer. More than 1000 IUCr articles have been indexed and included in ChemRefer's search engine as a result.]

Discussion. Questions and Answers

Jim Downing*:* Don't just think of your own use case. Open standards are important in allowing novel uses of data that weren't originally designed in.
Jeremy Frey (University of Southampton): There are problems in keeping equipment working. Research workers have some flexibility and can use the equipment when it is available but in teaching the equipment has to be available at a certain time.

Colin Batchelor (RSC): Jabber (http://www.jabber.org/) can incorporate CML in instant messaging. Jabber is an instant messaging protocol that supports XML, and therefore CML for exchanging chemical data. It's what Google Talk uses to implement instant messaging. [Postscript: the protocol has become a standard and is now called XMPP.]
Gráinne Conole: Wikis and blogs are not much use for rich semantic data.
Jeremy Frey: It depends on whose wiki or blog you are using.
Leslie Carr (University of Southampton): Web 2.0 has potential. Univillage (http://www.univillage.com) is a social networking site aimed UK university students (similar to Facebook (http://www.facebook.com/) in the United States). In my teaching in computer science, I found that everyone signed on. Is this wishful thinking or "new toyism"?
Gráinne Conole: We don't know yet. LXP shocked me. The true implications are only just starting to appear. The students are doing all these things, even the non-techies.
Unknown: Don't put chemical information into an institutional repository; put it in YouTube.
Gráinne Conole: Whatever is done, it must be done quickly.
Colin Batchelor: Are you looking at how undergraduates are using the technology?
Gráinne Conole: There is not enough money. This is a tiny project in four subject areas. It needs extending. Oxford Brookes University is reviewing this area. We need to do much more.
Unknown: Put the information in an institutional repository but make the interface simple enough that it can be integrated. Our interfaces are not simple enough. Distribution at institutional level works against us. Subject-level distribution might be better.
Jeremy Frey: We need to do both. We need to get out there quickly. Data without provenance can be captured. You can link data and keep the provenance but you do need to capture the provenance otherwise the data might be interesting but you could not trust them.
Rachel Heery (UKOLN): [Question addressed to Peter Strickland.] Who does the links to other databases?
Peter Strickland: It depends on the link. If the link is to PDB, you need to check if the data are in PDB, so IUCr checks that, but with CCDC there is an automatic look-up. There are two main places where we would make links to data in our articles. Links at the article level (i.e., the data set is strongly related to the current article) would apply to the links we make to the PDB. The other case is links at the level of a citation in the reference list. We can do a bibliographic search on a structural database and link a reference to the data corresponding to that paper. This applies, for example, to the links we make to the CCDC.
Unknown: Where is the added value created?

Breakout Sessions

In the afternoon the attendees separated into three groups, each group to consider just one aspect of Phase 3 feedback. Each group was asked to make five crisp recommendations, applicable specifically to the crystallography domain. The subject headings for the three groups were laboratory data, repository interoperability, and the partner landscape.

Laboratory Data

Recommendations

- Capture context at the beginning of the process to minimise information loss.
- Researchers should be encouraged to have a plan prior to doing an experiment. They can pull in data from outside to add context to the plan. This makes it easier to capture workflow. Use instrumentation to capture the data.
- Note that there are two views or information streams: that from the instrument and that from the laboratory environment. Some information is recorded in the electronic laboratory (e-lab) notebook but instruments and environments have their own logs that automatically capture status of the equipment etc. Details of a diffractometer experiment would be correlated by time and place to get equipment information and an e-lab log, but users interested only in the results of the experiment would not need this level of detail.

- The strategy for access rights and control needs to be established at the very beginning and linked to planning. Researchers need information only about their own samples; technicians need access to more information.
- Educate people in methodologies for recording data fully with context. These users would immediately see the advantages in terms of data reuse by the individual who recorded the data or by other users.
- Record information with a schema and establish a place to put these schemas so that others can look at them.
- Ensure the security of the repository. Security features must be built-in, not added on afterwards.

Detailed discussion

The eBank project came out of the CombeChem eScience project. Data are at the heart of the project and data capture starts in the laboratory. We need efficient and future-proof capture in the laboratory, so we need to capture context at source, and then to archive the information. It is essential to capture *digital information*. The current ELN (electronic laboratory notebook) software is on a pervasive network, where a plan may be created, viewed, annotated or actioned from anywhere, e.g., the office, a tablet PC in the laboratory, or a USB balance. In CombeChem, it was recognised that chemists have to consider Control of Substances Hazardous to Health (COSHH) regulations when they plan experiments, so the ELN plan was COSHH-driven, not built on a global model. Central facilities operate safety planning and proposals but this is still a form of plan. The scientist would write a scientific case, and identify materials to be examined, and the experimental conditions required, and a plan could be deduced or inferred.

How many needless experiments are performed due to the lack of planning? COSHH regulations were initially seen as a hindrance but they are now applied in normal practice. Unfortunately they are difficult to apply to all areas of chemistry. Will they be so much of a barrier that there will be no take up? Hindsight is a wonderful thing. It all comes back to teaching. Currently, chemists are learning on the job from their peers: this is not best practice. It is important that they be taught about planning, and the value of information gathering, dissemination and contextualisation, and the benefits of looking at the bigger picture.

Many “digital” experiments, for example in physical chemistry, are carried out on instruments, and it is possible to “log” them. Larger instruments with big data streams are a good place to start training because, in this case, everything falls down if there is no management. Some logging technology is supplied on modern instruments but it uses proprietary software. Manufacturers are keen to develop logging software.

Experimental schemas have to be considered. There is a need to bring the synthesis record into a crystallography experiment. (A synthesis record is a recipe with annotation as to how it was followed, which aids the analysis of the product.) Currently this information is not provided. There is a need to record the software, and software version used. The instrument and operational parameters must be captured. An independent instrument log with instrument status and configuration must be added to the repository record. There are two data streams which are independent of each other, the laboratory environment and the experiment. To make a complete experiment log, the environment data must be bought into CIF or NeXus. This can be done on the fly by requesting the environment data for the time period in question. (NeXus is a dictionary-based interchange format like CIF, and indeed contains some of the information that CIF does, but adds more information and detail on the experiment and is capable of handling many different types of experiment i.e., not just crystallography.) Is there a need for the sample preparation log in a microscope operation repository? A “dark” repository or escrow may be required for samples submitted by commercial organisations; the security model would identify access rights.

Is good e-laboratory practice good enough for curation? The approach to curation is a “whole life cycle” one. “Open source” software supplied by academics (available from a Web site) causes

problems with undocumented updates and versions. The file that was presented to the software and the resulting output should be kept for reproducibility. Perhaps a software archive is needed. Workflow capture is needed for reuse, reproduction, and peer review of information. Authoring tools could be supplied for papers subject to good e-laboratory practice standards. The workflow requires annotation if it is to be re-used. People must be able to dip in and out of the workflow. Multiple people may be involved in the one experiment and they may not all have prior information.

Who will pay for sustainability of research outputs? Anything is better than nothing if there is no finance, but data reuse will be very limited. Finance will affect the lifetime of the data. Raw data might have a finite lifetime if all the requisite information has been extracted from them but results data are part of the life cycle and must be kept. Institutions making institutional repositories give no guarantees on preservation of the data. Curation experts and learned societies have an overlapping role in archiving and preservation.

Repository Interoperability

Recommendations

- Re-pose the same questions that eBank Phases 1-2 posed but pose them to a wider group of partners in order to benefit from more recent experience and get support for existing solutions, or extended ones.
- We do not yet know what the common services will be but we agree we should participate.
- We should agree on the process of making technical decisions.
- We should revisit all technical issues.
- We should agree on the roles of the stakeholders and establish what contributions the partners are prepared to make.
- All the old chestnuts of identification (Uniform Resource Identifiers (URIs), or URLs, or DOIs) need to be re-addressed.
- Consider if the decision to use DOIs is right for a larger federation. Publishers might recommend use of the DOI but in wider partner communities that might not be so appropriate.
- We need to be sure what we are putting into our repositories. Consider the following questions. Will the repositories extend beyond crystallography? Have we got our data model right? When we resolve a DOI, do we know what we are getting?
- Maintain the integrity of the system as a whole (e.g., do not make random *ad hoc* instances and give new identifiers etc.) when addressing workflow complexity and all the different processes etc. involved.
- Define metadata application profiles: how the objects are described. Agree on content packaging, i.e., the representations of objects we want to pass around. Agree on APIs and the range of functionality that the federation will support.
- Maintain awareness of similar discussions with the Object Reuse and Exchange (ORE) project.
- Ensure that the federation is well represented in ORE. There is a need to join up with these developments in the United States; a need for grounding in good infrastructure.

Detailed discussion

The DOI Handle question needs to be re-addressed. Also we must automate registration. eBank chose to use DOIs for data sets. There are Simple Object Access Protocol (SOAP) delivery issues. There are seven DOI agencies and two for “data sets”; one was affordable. The Technische Informationsbibliothek und Universitätsbibliothek Hannover (TIB) solution should be revisited and compared with the CrossRef, IUCr, United Kingdom Education and Research Network Association (UKERNA) and JISC solutions. It is proposed that the federation require

every public identifier to be a URI, guaranteed by its institution. We need a shared model of the referent, a model of what we are dealing with.

Investigate issues of workflow complexity. Exchange data and identifiers whilst maintaining integrity. Create copies with different URIs. Ensure the integrity of URI assignments.

Is the information to be stored in a subject-based repository or an institutional repository? Define metadata application profiles.

What representations of the objects should be shared between our systems? Once this is resolved, consider content packaging according to the METS standard.

Agree on APIs and exposed functionality, including deposit.

The federation has not yet agreed on common, federated services. Group Two recommended, based on the experience and understanding gained in eBank Phases 1 and 2 that we:

- agree how to make technical decisions
- revisit all technical issues, and
- agree on the roles of stakeholders and on partner contributions.

Partner Landscape

Recommendations

- Federation, if standardised, can produce a trusted teaching tool suitable for the “digital natives” (students born into the era of digital technologies and multitasking). Fast adoption of new technologies by students is in contrast to the slowness of changes in teaching.
- Diffusion of the learning aspect may be helped by open repositories and hindered by closed ones.
- The federation should focus on curation of data i.e., long term sustainability and data preservation, despite the problems of funding and indecision about who owns the problem.
- Agreement on standards is needed. Publishers have adopted (not imposed?) certain standards. IUCr felt that standards arise to meet community needs. Spectra are different from crystal structures because instrument manufacturers who have used their own commercial standards have to be brought on board.
- The federation needs to look at standards for spectra and perhaps take a broader outlook, rather than concentrating only on crystallography
- The federation needs high quality data, even though it is possible to get good conclusions out of a poor crystal. There is a need to capture metadata at source.
- Consider the establishment of a trusted repository and how to measure “trust”.
- Note that the National Science Foundation (NSF) has recommended programmes that help automation of instrumentation.

Detailed discussion

The initial discussion centred on non-technical, socio-political issues. One delegate said he was pleasantly surprised at the cheminformatics course results presented by Gráinne Conole: it was far more structured than he expected. This could be related to the quality of the data or the quality of the students. Students tend to be critical but they were not so this time. A delegate wondered about the contents of the course and another asked if Southampton intended to build on this experience. It seems that there has been some support for the idea. A resource is needed for such a course. This fits in with the repository concept. The Southampton resource is on Blackboard behind the Southampton firewall. It was observed that changes in teaching practice

happen slowly and making a resource available would at least remove an inhibitor. It was confirmed that records could be made available from the crystallography archive and that many concurrent Cambridge Structural Database (CSD) users could be accommodated by the Chemical Database Service.

There was discussion about letting students get “hands-on” and about quality control. The key is federation: Southampton has some data but this is still not federation. It is *federation* that matters. Federation implies consistency. A subset of data files needs devising, to established standards, so that the content can be trusted by teachers. A distributed model would be used for the information. Federation could mean that everyone contributed to one central source but this is not necessarily important. A delegate asked if the contributions were data or learning materials. Another said that the obligation of the federation is to be able to understand a question and answer it in the same way. Perhaps data for teaching need separating out, i.e., a subset is needed.

The original idea of an institutional repository was that it held data belonging to that institution. No university has agreed on curation for ever and ever. One idea of federation is sharing so that preservation is guaranteed. Whether the data are held in only one place may depend on costs and local policy. Are data centres in a position to make a long term commitment? Some people think that this is a problem for the UK research councils but it is an international problem too. Everything has a cost. RCUK supported open data but they must put their money where their mouths are. In the interests of long term curation, the IUCr might be interested in mirroring anything that were made available but IUCr cannot do long term archiving if the data are all over the place and in many different formats. Standards must be mandatory. The core information will be in CIF or XML.

A delegate claimed that publishers have had an effect on the community by imposing critical standards. They can refuse to publish unless the material is in the right format. A publisher disagreed with this but agreed with IUCr that publishers work with repositories to establish a standard, i.e., standards arise to meet community needs. IUCr confirmed that it does not *impose* standards. Two crystallographers emphasised that crystallography *is* standardised. The concept of publishers imposing standards is perpetuated by Peter Murray-Rust who talks about “destruction by publication” and says that your work will not be published unless you conform.

A discussion followed on standards for storing and retrieving spectral data and whether the federation should look more widely than just at crystallography. A delegate reported attending an NSF meeting where a program to influence vendors of analytical instruments was discussed. Perhaps NSF and JISC should talk to each other about this. A pincer mechanism might influence both instrument manufacturers and publishers. It was stated that JISC has left sustainability hanging in the air. Someone suggested that software for conversion between different spectral data formats might help.

IUCr said that it is essential to have high quality data but another crystallographer pointed out that what is needed is the correct interpretation from the data that are available. A correct interpretation can actually be derived from some pretty awful data. The quality of the sample determines the quality of the data and it is possible to get valuable data from a bad sample. It is important to preserve the images. The eCrystals model has all the data that is needed: raw data, the CIF and links are archived.

It is important to record the provenance of data. If you keep track of the steps you will not have to repeat all the analysis. In satellite image interpretation dozens of steps need contracting. It is no different in chemistry. There is nothing in the metadata in a crystallography experiment that makes quality obvious but the information is in the supplementary files. The validation program checkCIF is peculiar to crystallography. In spectral data, assignments are made. These are perhaps slightly subjective, unless you do a quantum mechanical calculation. The signal to noise

ratio matters in some fields. So the concepts of quality and provenance are not quite the same in spectra as in crystallography.

No-one was able to give an update on the International Spectroscopic Data bank (IS-DB) project led by Tony Davies. A crystallographer noted that his instrument captures even more data than he can use but capturing metadata off analytical instruments should be possible. If an archive were compulsory, there would be an incentive to capture the metadata. The RSC expressed a willingness to handle a spreadsheet (which one delegate hoped would not be in Excel format). We should keep to a minimum what needs to be captured: "the minimum set of data items with the maximum possibilities".

Trustedness of institutional repositories was next discussed. Trustedness can be assessed by applying questions and criteria. It is not clear if this will work or what it will cost. It is not possible to assign a nine out of ten type of score.

Final Discussion

Liz Lyon: Do the partners want to add comments about the schema?
Alan Tonge: This is an important area: not just the crystallography metadata schema. Crystallography differs from most areas. The crystallography issues have been sorted out.
Peter Morgan (Project Director, SPECTRa, Cambridge University): We are committed to this already.
Liz Lyon: I will rephrase the question. *How* should we go forward with the metadata model and the tools in the laboratory?
Rachel Heery: Which data are we talking about? Not just crystallography? Some *policy* decisions are needed.
Alan Tonge: Should there be a "working party"?
Liz Lyon: At the Southampton-UKOLN meeting yesterday we thought of going to the partners with semi-structured questionnaires but now I am wondering if that is the right way ahead.
Chris Rusbridge (Director, DCC): eBank cannot impose standards so we must revisit the standards, but how do we do this.
Simon Coles: The straw man.
Liz Lyon: We need to consider the laboratory as well as the repository. [At this point, others in the audience pointed out that instrument manufacturers are "secretive".] This will be challenging.
Jeremy Frey: Instrument manufacturers consider that it is not worth further developing their software once they have gained competitive value from it.
Chris Rusbridge/Simon Coles: But there is a standard "ticklist".
Jeremy Frey: Analytical Information Markup Language (AnIML**)** is the developing XML standard for analytical chemistry data.
Mike Hursthouse (University of Southampton): What are the technical issues that need revisiting? I thought that DOI was the only standard at issue. All that remains to be decided is long term preservation.
Simon Coles: We can't impose eBank standards on everyone.
Mike Hursthouse: eBank is a working model.
Chris Rusbridge: Your working model might not have the critical mass of acceptance. You have to be humble about it.
Rachel Heery: I return to my earlier point. What is the federation trying to do? Just grow an IUCr collaboratory or grow wider?
Mike Hursthouse: I thought that "massage" not "revisiting" was needed.
Leslie Carr: We don't want to invalidate or change the decisions made in the crystallography community but we want to move beyond the confines of this community. JISC wants to see solutions that can be adapted more widely but we musn't be seen to be *undoing* what we have done.
Simon Coles: eBank was a demonstrator for applications in other areas.

Mike Hursthouse: “This is how we did it. You can do it this way if you like”. I was shocked by Les Carr’s flipcharts [i.e., the feedback from the second breakout group]. Repositories *do* need to address longevity, reproducibility/replication, and security. Don’t mix up R4L with eCrystals.
Simon Coles: We have a robust system for crystallography in our laboratory but we want to look wider and put together all the pieces of the jigsaw.

Glossary

ALPSP	Association of Learned and Professional Society Publishers	http://www.alpsp.org
AnIML	Analytical Information Markup Language	http://animl.sourceforge.net/
API	Application Programming Interface	
ARROW	Australian Research Repositories Online to the World	http://www.arrow.edu.au/
Blackboard	A commercial virtual learning environment	http://www.blackboard.com/
Blog	A website where entries are made in journal style and displayed in a reverse chronological order	
CAS	Chemical Abstracts Service, a division of the American Chemical Society	http://www.cas.org
CASPAR	Cultural, Artistic and Scientific knowledge for Preservation, Access and Retrieval	http://www.casparpreserves.eu
CASRN	Chemical Abstracts Service Registry Number, a unique numeric identifier designating only one substance	http://www.cas.org/EO/regsys.html
CBF	Crystallographic Binary File	http://www.iucr.org/iucr-top/cif/imgcif/index.html
CCDC	Cambridge Crystallographic Data Centre	http://www.ccdc.cam.ac.uk/

CCLRC	Council for the Central Laboratory of the Research Councils	http://www.cclrc.ac.uk/
CETL	Centres for Excellence in Teaching and Learning	http://www.hefce.ac.uk/Learning/tinits/cetl/
checkCIF	A Web service that reports on the consistency and integrity of crystal structure determinations reported in CIF format	http://checkcif.iucr.org/
Chemistry Central	An emerging Open Access series of journals	http://www.chemistrycentral.com
CIF	Crystallographic Information File and Crystallographic Information Framework (a broader system of exchange protocols)	http://www.iucr.org/iucr-top/cif/home.html
CLADDIER	Citation, location, and deposition in discipline and institutional repositories	http://claddier.badc.ac.uk/trac
CML	Chemical Markup Language. Applies XML to the management of chemical information	http://cml.sourceforge.net/
COD	Crystallography Open Database	http://www.crystallography.net
CODATA	ICSU Committee on Data for Science and Technology	http://www.codata.org/
CombeChem	An eScience project involving the “end-to-end”’ linking of data and information	http://www.combechem.org/

COMCIFS	IUCr Committee on the Maintenance of the CIF Standard	http://www.iucr.org/iucr-top/cif/index.html#comcifs
COSHH	Control of Substances Hazardous to Health	http://www.coshh-essentials.org.uk/
CrossRef	An independent membership association, founded and directed by publishers. It is the official DOI link registration agency for scholarly and professional publications	http://www.crossref.org
CrystalGrid Collaboratory	A US-based consortium of partners (also including the University of Sydney, Australia and NCS) sharing and publishing crystallographic data by means of Open Access data repositories based at each of 20 sites	http://www.crystalgrid.org/ **[*Domain name required re-registration in December 2006]** http://eprints.soton.ac.uk/9777/
CRYSTMET	Metals structure database. A structure and powder database for metals and intermetallic compounds	http://www.tothcanada.com/databases.htm
CSD	Cambridge Structural Database	http://www.ccdc.cam.ac.uk/products/csd/
DAREnet	Digital Academic Repositories	http://www.darenet.nl/en/page/language.view/search.page
DCC	Digital Curation Centre	http://www.dcc.ac.uk/

Deposit API	An agreed technical description of the transaction that happens when an item is deposited in a repository	http://www.ukoln.ac.uk/repositories/digirep/index/Deposit_API
DOI	Digital Object Identifier	http://www.doi.org/
DSpace	A digital repository system that captures, stores, indexes, preserves, and distributes digital research material	http://www.dspace.org/
Dublin Core	The Dublin Core metadata element set is a standard for cross-domain information resource description	http://dublincore.org/
eBank UK (eBank for short)		http://www.ukoln.ac.uk/projects/ebank-uk/
eCrystals	An institutional repository that makes available the raw and derived data from a crystallographic experiment	http://ecrystals.chem.soton.ac.uk http://eprints.soton.ac.uk/41257/
ELN	Electronic Laboratory Notebook	
eMalaria	An integrated Web-based system, for use in schools, for the design and testing of small drug-like molecules against an enzyme of known structure	http://chemtools.chem.soton.ac.uk/projects/emalaria/
ePrints	Software for managing an institutional repository	http://www.eprints.org/

ePrints Soton	The University of Southampton's research repository	http://eprints.soton.ac.uk/
Folksonomy	A folksonomy is an Internet-based information retrieval methodology consisting of collaboratively generated, open-ended labels (or "tags") that categorise content. It differs from a taxonomy in that the authors of the labelling system are often the main users (and sometimes originators) of the content to which the tagging is applied.	http://en.wikipedia.org/wiki/Folksonomy
GRADE	Geospatial Repository for Academic Deposit and Extraction	http://edina.ac.uk/projects/grade/
Handle	The Handle System is the resolution component of the DOI System	http://www.doi.org/handbook_2000/glossary.html
HTTP	Hypertext Transfer Protocol	http://www.w3.org/Protocols/
ICDD	The International Centre for Diffraction Data	http://www.icdd.com/
ICSD	Inorganic Crystal Structure Database	http://icsd.ill.fr/icsd/index.html
ICSTI	International Council for Scientific and Technical Information	http://www.icsti.org

ICSU	International Council for Science, (formerly the International Council of Scientific Unions)	http://www.icsu.org
ICT	Information and communication technology	
imgCIF	A CIF dictionary of data names required by the Crystallographic Binary File (CBF) image representation project	http://www.iucr.org/iucr-top/cif/imgcif/index.html
InChI	IUPAC International Chemical Identifier	http://www.iupac.org/inchi
Intute	The new name for the Resource Discovery Nework (RDN) service. Intute is a composite word derived from “Internet” and “Tutorial”	http://www.intute.ac.uk
IS-DB	International Spectroscopic Data Bank	http://www.is-db.org/
ISIS	The pulsed neutron and muon source situated at the CCLRC Rutherford Appleton Laboratory. ISIS is not an acronym	http://www.isis.rl.ac.uk/ http://www.isis.rl.ac.uk/aboutIsis/index.htm
IUCr	International Union of Crystallography	http://www.iucr.org/
IUPAC	International Union of Pure and Applied Chemistry	http://www.iupac.org
JCAMP	Joint Committee on Atomic and Molecular Physical Data	http://www.iupac.org/standing/cpep/wp_jcamp_dx.html

JCAMP-DX	In 1995 IUPAC took over responsibility for the range of scientific standards for exchange of spectral data from the Joint Committee on Atomic and Molecular Physical data and the group Data eXchange	http://www.iupac.org/standing/cpep/wp_jcamp_dx.html
JISC	Joint Information Systems Committee	http://www.intute.ac.uk/sciences/
JISC Deposit API	See Deposit API	http://www.ukoln.ac.uk/repositories/digirep/index/Deposit_API
LEX	The learner experience of e-learning project	http://www.jisc.ac.uk/whatwedo/programmes/elearning_pedagogy/elp_lex.aspx
LXP	Learner experiences of e-learning	http://www.jisc.ac.uk/media/documents/lxp_project_final_report_nov_06.pdf
Mashup	A Web site or Web application that combines content from more than one source	http://en.wikipedia.org/wiki/Mashup
METS	Metadata Encoding and Transmission Standard. The METS schema is a standard for encoding descriptive, administrative, and structural metadata regarding objects within a digital library, expressed using the XML schema language of the World Wide Web Consortium.	http://www.loc.gov/standards/mets/
mmCIF	Macromolecular Crystallographic Information File	http://www.iucr.org/iucr-top/cif/mm/index.html

Molfile (MOL)	A file format for representation of chemical structure information	http://www.mdl.com/support/knowledgebase/faqs/faq_ib_27.jsp
NCeSS	National Centre for e-Social Science	http://www.ncess.ac.uk/
NCS	UK National Crystallography Service	http://www.soton.ac.uk/~xservice/
NeXus	A common data format for neutron, X-ray, and muon science.The archiivng format used by ISIS (and others)	http://www.nexusformat.org/Main_Page
OAI	Open Archives Initiative	http://www.openarchives.org/
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting	http://www.openarchives.org/OAI/openarchivesprotocol.html
OAIster	A metadata harvester that provides a generic retrieval resource for information about publicly available digital library resources	http://oaister.umdl.umich.edu/o/oaister/
Ontology	A data model that represents a domain and is used to reason about the objects in that domain and the relations between them	See, for example, http://www.w3.org/2001/sw/WebOnt/
OpenURL	A standard syntax to create Web-transportable packages of metadata and identifiers about an information object	http://www.niso.org/committees/committee_ax.html
ORE	Object Reuse and Exchange	http://www.openarchives.org/ore/

PDB	Protein Data Bank	http://www.rcsb.org/pdb/home/home.do
Podcast	A multimedia file that is distributed by subscription (paid or unpaid) over the Internet using syndication feeds, for playback on mobile devices and personal computers	
Psigate	Now the Science, Engineering and Technology component of Intute	http://www.intute.ac.uk/sciences/
R4L	Repository for the Laboratory	http://r4l.eprints.org/about.html http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_r4l.aspx
RCUK	Research Councils UK, the strategic partnership of the UK's eight Research Councils	http://www.rcuk.ac.uk/
Reciprocal Net	A distributed database used by crystallographers to store information about molecular structures	http://www.reciprocalnet.org/
RSC	Royal Society of Chemistry	http://www.rsc.org/
Semantic Web	A common framework that allows data to be shared and re-used across application, enterprise, and community boundaries	http://www.w3.org/2001/sw/

Smart Tea	An electronic laboratory notebook project focusing on the experimental design and execution process. Part of the CombeChem project	http://www.smarttea.org/
SOAP	Simple Object Access Protocol	http://www.w3.org/TR/soap/
Social tagging	See Folksonomy	
SPECTRa	Submission, Preservation and Exposure of Chemistry Teaching and Research Data. A Digital Repository for the Chemical Community	http://www.lib.cam.ac.uk/spectra/
SRW	An XML-based protocol designed to be a low-barrier-to-entry solution for information retrieval operations across the Internet, using Common Query Language (CQL)	http://www.loc.gov/standards/sru/srw/
StORe	Source-to-Output Repositories	http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_store.aspx
TIB	Technische Informationsbibliothek und Universitätsbibliothek Hannover	http://www.tib.uni-hannover.de/
UKERNA	United Kingdom Education and Research Network Association. UKERNA manages the operation and development of the JANET education and research network	http://www.ja.net/about/index.html

UKOLN	Formerly “UK Office for Library Networking”	http://www.ukoln.ac.uk/
URI	Uniform Resource Identifier	http://www.w3.org/Addressing/URL/uri-spec.html
URL	Uniform Resource Locator	http://www.w3.org/Addressing/URL/Overview.html
WebDAV	Web-based Distributed Authoring and Versioning. A set of extensions to the HTTP protocol which allows users to edit and manage files on remote web servers collaboratively	http://www.webdav.org/
Weblog	See blog	
Wiki	A type of Web site that allows the visitors themselves to edit and change available content easily, sometimes without the need for registration. An effective tool for collaborative authoring	
XML	Extensible Markup Language	http://www.w3.org/XML/
XMPP	Extensible Messaging and Presence Protocol	http://www.xmpp.org/
Z39.50	An international standard for communication between computer systems, primarily library and information related systems	http://www.loc.gov/z3950/agency/