Scaling Up: Towards a Federation of Crystallography Data Repositories

Document details

Author:	Liz Lyon, Simon Coles, Monica Duke, Traugott Koch
Date:	12th May 2008
Version:	1.0 Final
Document Name:	ebank-phase3-report-final.doc
Notes:	

Acknowledgement to contributors

The authors would like to thank the various people, who contributed to the report by completing an interview or commenting on previous versions. The authors take responsibility for interpreting the answers and for any change of emphasis that comes with collating the viewpoints of the various contributors.

Acknowledgement to funders

This work was funded by the JISC as part of the Digital Repositories Programme.

UKOLN is funded by the MLA: The Museums, Libraries and Archives Council, the Joint Information Systems Committee (JISC) of the Higher and Further Education Funding Councils, as well as by project funding from the JISC and the European Union. UKOLN also receives support from the University of Bath where it is based.

1 Executive summary 5

2 Introduction 7

2.1 Background and Context 7

2.2 Audiences 8

3 Methodology 8

3.1 Desk-based research 8

3.2 Consultation Workshop 10

3.3 Interviews 10

4 Findings 10

4.1 Cambridge Crystallographic Data Centre (CCDC) 10

4.2 Chemical Database Service (CDS) 13

4.3 Chemistry Central 14

4.4 International Union of Crystallography (IUCr) 15

4.5 ReciprocalNet, Indiana University 16

4.6 Royal Society of Chemistry (RSC) 17

4.7 Science & Technology Facilities Council (STFC) 19

4.8 University of Cambridge / SPECTRa-T Project 22

4.9 University of Sydney, Australia 25

5 Synthesis and Discussion 26

5.1 Institutional Repositories Policy and Practice 26

5.2 Crystallography Laboratory Practice and Workflows 28

5.3 Technical Interoperability and Standards 30

5.4 Metadata Schema and Application Profiles 31

5.5 Semantic Interoperability 32

5.6 Data Citation, Identifiers and Linking 33

5.7 Federation Architectures and Third Party Services 34

5.7.1 Levels of Service 34

5.7.2 Solutions and experience from the digital library community. 36

5.7.3 Interactions with third party services 37

5.7.4 Evaluation of architectural options. 40

5.8 Rights and Licensing 41

5.9 Data Quality and Validation 42

5.10 Preservation, Curation and Sustainability 42

5.11 Community and Inter-disciplinary Interactions 44

5.12 Collective Intelligence and Open Science 45

6 Appendix 46

6.1 Interview Pro-forma 46

6.2 List of individuals participating in the interviews. 48

7 References 49

1 Executive summary

The *Scaling Up Report* presents the results of a JISC-funded scoping study to assess the feasibility of a federated model for data repositories in the domain of crystallography. It builds on earlier work in the eBank UK Project and has been based on a mix of desk-based research, a consultation workshop and a series of interviews with stakeholders.

The Synthesis is presented in twelve sections: Institutional Repositories Policy and Practice, Crystallography Laboratory Practice and Workflows, Technical Interoperability and Standards, Metadata Schema and Application Profiles, Semantic Interoperability, Data Citation, Identifiers and Linking, Federation Architecture and Third Party Services, Rights and Licensing, Data Quality and Validation, Preservation, Curation and Sustainability, Community and Inter-disciplinary Interactions, Collective Intelligence and Open Science.

The authors conclude that a federation-based approach is an appropriate strategy for this domain and a Checklist of Community Criteria for Interoperability, summarises the elements which contribute a solid foundation for the model.

Community Criteria for Interoperability	Crystallography exemplars
1. Involvement of professional bodies and publishers.	Royal Society of Chemistry, IUCr.
2. Development and adoption of a common domain data format standard.	CIF
3. An established data validation mechanism.	CheckCIF.
4. Implementation and adoption of a common domain identifier.	InChI
5. A metadata schema application profile which supplies a common core element set.	eBank-UK schema
6. An existing subject repository, which may operate on a commercial basis.	CCDC
7. A degree of homogeneity and co-ordination in disciplinary research practice.	CIF and COMCIFS
8. An established service ethic and associated policies, which drives research practice for the common good.	NCS or CCDC or CDS

Conversely a number of Disruptive Effects act as constraints and barriers, and inhibit inter-disciplinary interactions.

Disruptive Effects	Mitigating Action
1. Diversity of internal laboratory practice and culture.	Best practice standards, advocacy, core standard formats, AP
2. Arbitrary re-use of data because of “lock-in” to instrumentation and proprietary software e.g. CSD.	Advocacy, core standard formats, AP

3. Data re-use is limited because only processed (not raw) data is shared more widely.	Capture and expose raw data in laboratory repositories.
4. Limited data-sharing culture within crystallography, which inhibits wider chem-informatics.	Advocacy, awareness-raising, tool development
5. Inter-disciplinary re-use of data depends largely on human interaction and is hindered by lack of m2m interfaces.	Develop Web services such as CrystalEye which operate across distributed repositories.
6. Formal publishing disconnects inhibit interdisciplinary interactions e.g. lack of embedded links between domain identifiers such as LSIDs and InChIs.	Advocacy, awareness-raising, and partnerships with publishers. Develop knowledge extraction tools
7. Competitive relationships between institutions, departments and laboratories, as a result of research assessment frameworks and funding awards.	Consortium agreements should include clauses on data-sharing.
8. High-level strategic fragmentation associated with data management plans within and between the funding bodies.	Co-ordinated strategic planning for data curation *across* research councils, other funders.

In addition, a number of Recommendations are made for further investigation.

- Recommendation 1: JISC should provide guidance to support the development, interoperability and sustainability of sub-institutional repositories such as those at departmental, research group and laboratory levels.
- Recommendation 2: JISC should consider funding an investigation of “laboratory informatics” including LIMS, to identify opportunities for more generic workflow integration and pervasive systems to capture laboratory data and metadata in-situ.
- Recommendation 3: JISC should support further work to explore alternative and/or automatic assignment of terms and keywords to data sets for enhanced discovery.
- Recommendation 4: JISC should seek expert advice to advocate the implementation of appropriate open data licences to provide a common basis for data sharing within the research community.
- Recommendation 5: JISC should consider funding further work to support data validation and data quality assurance methodologies, possibly taking a domain-centric approach.
- Recommendation 6: JISC should fund the development of quantitative criteria for the appraisal of datasets. These criteria should take into account how the reproducibility of an experiment can be described in a “standard” manner.
- Recommendation 7: JISC should fund a scoping study to investigate the potential of collaborative technologies, collective intelligence and repository content and services, to stimulate new modes of open science.

2 Introduction

This Report presents the outcomes of the eBank-UK Project Phase 3 Scoping Study, which investigated the feasibility of the proposed eCrystals Federation of data repositories. It is the main deliverable from this final Phase of the Project. The Report contains:

- A description of the methodologies used.
- The collated findings from the desk-based research and the interviews.
- Synthesis and discussion based on these findings, building on the outcomes from the earlier Consultation Workshop.
- Commentary on a range of Perspectives and Recommendations for further work.

2.1 Background and Context

The eBank-UK Project (JISC-funded in three phases since September 2003), has investigated the feasibility of data repositories for the archiving and storage of crystal structure data, and the linking from primary data to other research outputs within the scholarly knowledge cycle[1]. Building on the Open Archive Initiative (OAI) concept, the project focussed on the laboratory based experimental technique of 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 is uploaded into a data repository and additional metadata (chemical & bibliographic), to Dublin Core standards, is associated with the dataset. 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, it 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 supported by the crystallographic governing body - The International Union of Crystallography (IUCr). All crystallographic data conventionally published in journal articles is collected by the Cambridge Crystallographic Data Centre (CCDC) and made available as the Crystal Structure 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 considerable interest in adopting the eBank-UK model for the publication of primary scientific data in a manner, which may be cited and linked to a formal article.

The transitioning of eBank to a federated model positions this project as a domain exemplar for the field of crystallography. The aim of expanding the number of participating partners managing data repositories reflects the changing nature of research practice towards a data-intensive paradigm and the model may be applicable to other disciplines. There are practical implications for full implementation arising from varied workflows in increasingly “smart labs” with the researcher requiring the tools and services to facilitate “digital scholarship”. There is also reference to open science constructs, which are emerging. The eCrystals Federation would build on the highly successful SHERPA experience in creating a network of institutional ePrint repositories. This Report describes foundational work examining the feasibility of such as Federation of data repositories. Whilst it is positioned in the domain of crystallography, the lessons learnt provide some generic guidance for other disciplines where the open publication of data is under consideration.

2.2 Audiences

It is hoped that a number of audiences will find this Report to be of interest, however the document is primarily aimed at funders and policy makers with an interest in promoting data publication, publishers and learned societies, institutional repository managers and research staff within the chemistry / crystallography discipline.

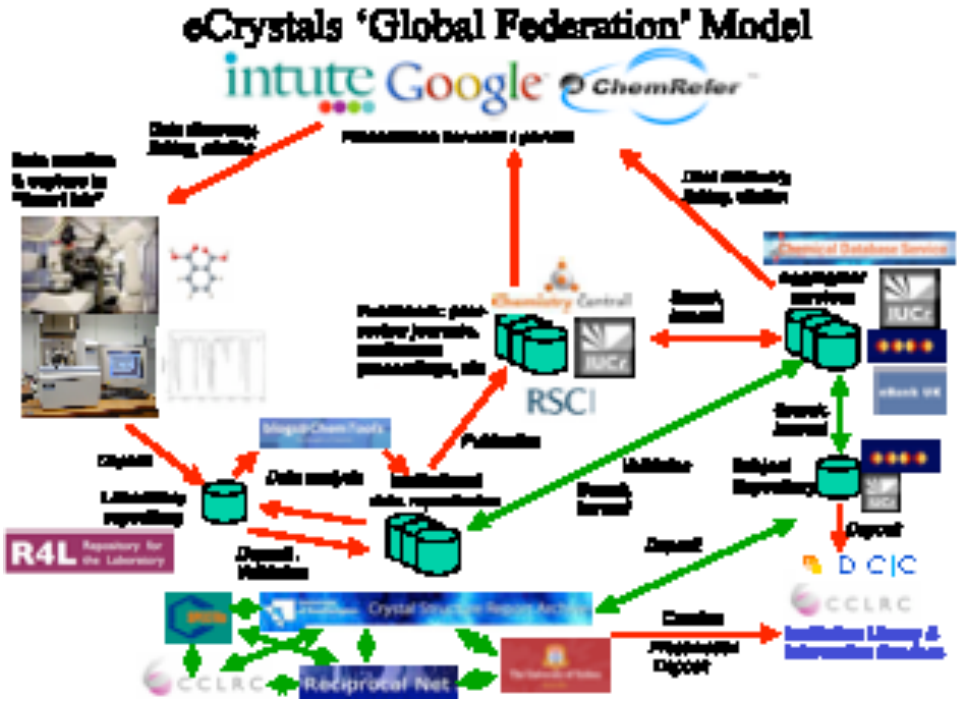
3 Methodology

The work has been largely based on two methodological approaches outlined below and supplemented by desk-based research. This report provides a synthesis of information and opinion gathered throughout the study with additional analysis and commentary.

3.1 Desk-based research

At the start of the Phase 3 work, a desk-based "snapshot" of the key stakeholders, potential partners and services in the Federation, was carried out. The aim of the exercise was to collect information that was easily available from the partner Web sites, thus identifying gaps in knowledge that could be explored during the interviews. A set of questions informed the information-gathering exercise. Partner organisations were identified from the project proposal and accompanying diagram. Figure 1 presents the Phase 3 "working diagram" with Federation entities, and was derived from the preliminary model included in the original Phase 3 proposal.

Figure 1 eCrystals Global Federation Model



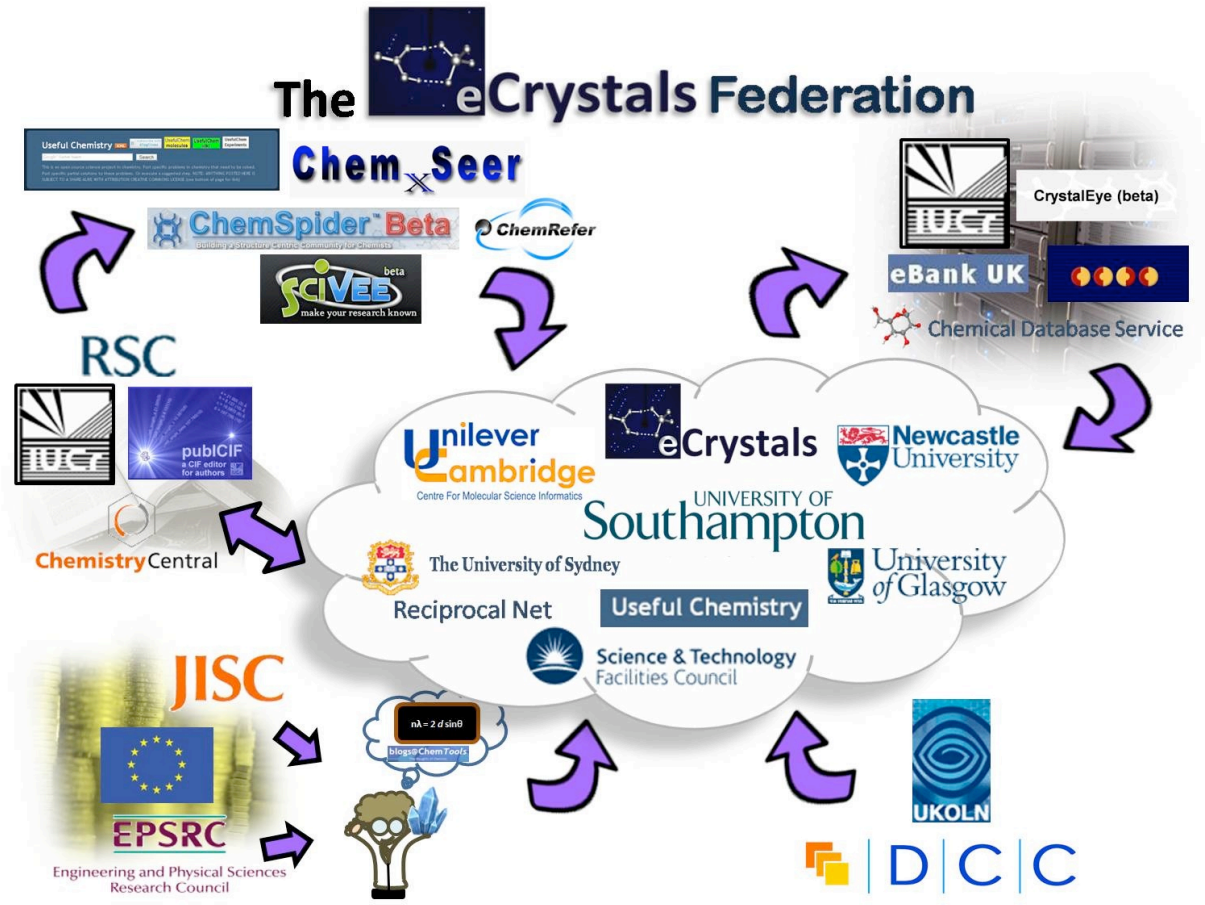


Figure 2 The eCrystals Federation Eco-system

Figure 2 represents the current view of the Federation: the list of entities includes partners, stakeholders and key services in the crystallography information environment:

- Southampton Crystal Structure Report Archive http://ecrystals.chem.soton.ac.uk/
- eBank UK aggregator service http://ebank.ukoln.ac.uk
- R4L Repository for the Laboratory http://r4l.eprints.org/
- DCC www.dcc.ac.uk
- Spectra http://www.lib.cam.ac.uk/spectra/
- STFC at RAL http://www.scitech.ac.uk/
- ReciprocalNet http://www.reciprocalnet.org/
- University of Sydney, Australia (includes http://mmsn.net.au/ Molecular and Materials Structure network)
- The International Union of Crystallography (IUCr) http://www.iucr.org/
- Royal Society of Chemistry (RSC) http://www.rsc.org/
- Chemistry Central http://www.chemistrycentral.com
- Chemical Database Service (CDS) http://www.cds.dl.ac.uk
- CCDC http://www.ccdc.cam.ac.uk/
- ChemRefer http://www.chemrefer.com/

- Intute http://www.intute.ac.uk
- Google http://www.google.co.uk

3.2 Consultation Workshop

This Workshop was intended to act as a bridge between the eBank Phase 2 and Phase 3 work. It was timed to provide an opportunity for the presentation of Phase 2 results, but also to create a forum where the prospective partners and stakeholders, could begin to identify and discuss the key issues to be addressed in the Federation model approach. The event was jointly supported and promoted by the three key data repository projects in the chemistry domain: eBank, R4L and SPECTRa.

Accordingly, an invitational workshop entitled "*Digital Repositories supporting eResearch: Exploring the eCrystals Federation Model"* was held at the Hilton London Metropole, London, on 20th October 2006. The purpose of the workshop was to:

- Develop a widespread understanding for the role of data repositories in scientific research, learning and dissemination.
- Scope an initial set of minimal requirements for a data repository to underpin the chemistry publication and dissemination processes
- Bring to light and probe issues surrounding interoperability, preservation, harvesting and aggregation in the data repository environment
- Produce an initial set of recommendations on schema design for construction of data repositories and data capture at the instrument level

The Workshop included a mix of presentations, breakout groups and discussions and allowed time for networking and collaboration. The remits of the breakout groups were

1) Capturing chemistry data in the lab: schema development, mechanisms for capture / ingest. Designing and managing data repositories: mechanisms for ingest, validation, presentation and OAI schema.
2) Federation and interoperability of repositories: OAI schema, interoperability standards, preservation and identifiers.
3) Learner, publisher, portal provider and data centre requirements in a repository enabled environment: linking to datasets from articles, division of content between article and repository, overlay journals, third party services and data repositories, pedagogic issues.

A full Report of the Workshop is available[2]. The *Scaling Up* Report seeks to build on, rather than duplicate, the contents of the earlier Workshop.

3.3 Interviews

A number of semi-structured interviews were subsequently held with selected representatives of the various stakeholder groups. The outline pro-forma used as a basis for the interviews is included in the Appendix together with the list of interviewees.

4 Findings

The findings are derived from a combination of Web-based information (factual profiles of non-core partners as shaded text boxes) and interview results based on the pro-forma (text reflecting opinion and views). Whilst every effort was taken to ensure consistency during the interview process through use of the pro-forma, the interviews varied to some degree in their composition and format, and deviations are noted in the text.

4.1 Cambridge Crystallographic Data Centre (CCDC)

Profile:

The Cambridge Crystallographic Data Centre (CCDC) is dedicated to the advancement of chemistry and crystallography for the public benefit through providing high-quality information services and software. CCDC operates the world repository for all crystallographic data published in journal articles, which comprises software for database access, structure visualisation and data analysis, and structural knowledge bases derived from this body of data. The CCDC serves the scientific community through the acquisition, evaluation, dissemination and use of the world's output of small molecule crystal structures by:

* Compiling the Cambridge Structural Database (CSD) - the world repository of small-molecule crystal structures

* Developing scientific products and services - structural knowledge bases and applications software for the life sciences and crystallography

* Maximising worldwide accessibility to the CSD for scientists in academia and industry

* Performing and supporting fundamental research using CSD information and CCDC products

* Promoting and supporting applications of crystal structure information in academia and industry.

The CCDC accepts depositions of crystal structure data from X-ray and neutron diffraction studies, and from powder studies using a constrained refinement, for organic and metal-organic compounds. Data depositions with the CCDC are of two main types:

* Pre-Publication: the structure(s) are being submitted for publication in a journal

* Private Communications to the CSD: the structure(s) are not intended for publication, but you wish them to be available to other scientists through the CSD

However, data for structures which have already appeared in a journal and are not yet in the CSD, are always welcome. The electronic CIF format should be used for all depositions. Since 1994, under official deposition arrangements with a number of journals, the Cambridge Crystallographic Data Centre (CCDC) has provided copies of the supplementary data of individual published structures for bona fide research purposes. Data from before 1994 are currently only available from the distributed Cambridge Structural Database (CSD). Supplementary data arriving at the CCDC electronically in CIF format, whether as part of journal deposition arrangements or directly from individuals, are held on trust in the CCDC Supplementary Data Archive on behalf of those journals and individuals. After publication, these data are converted into CSD entries by the addition of bibliographic and chemical text, chemical structural data, and the results of crystal structure validation.

Role in Federation: Supporting Partner, provides centralised crystal structures database.

The two primary issues that CCDC face are a) getting data into the Crystal Structure Database (CSD) and b) ensuring the accuracy of that data. Other difficult steps are making data publicly available and establishing the responsibilities for doing so. There is a further challenge in the "changing of mindsets" of many in the funding, research, university and publishing communities. Data publication timing issues, particularly between the CSD and journals were noted. One approach with data repositories is a time-delayed release, where the crystal structure data is made available at the point of publishing the paper. There were also issues around providing secure access to laboratory data archives for referees: views on this are divided with some people raising concerns, whilst others are not worried. The concept of publishing data not associated or allied to a publication i.e. independent publication rather than conventional publication, was of interest to CCDC, however it was noted that the American Chemical Society (ACS) is not in favour of this approach. The effect of independent data publication on the UK Research Excellence Framework (REF) is a prime concern. It was observed that publication through the *Acta Crystallographica E* route is one way to get data into the CSD, but there is also the 'Private Communication' route.

Another concern is clarity of ownership of data and this is viewed as being vitally important. It was stated that there is a need to categorise roles such as that of “creator”, and to allocate public responsibility for creation of a record. This aspect is particularly relevant to the eCrystals Federation partner organisations i.e. accountability versus responsibility in terms of datasets deposited in a repository. CCDC have data in a pre-publication archive but it cannot be made publicly available because permissions may not be given due to some or all of the associated people having “disappeared” i.e. individual contacts have been lost. CCDC has embargo systems in place, but has not so far made contact with authors. CCDC has a huge problem associated with making contact with these people: if the person has left an institution, then the email goes into a dead mail box and is lost.

There was some discussion of the maintenance of schemas and application profiles. The CCDC want to harvest data from repositories, and to maximise the amount of unpublished data coming through from these repositories. In cases where there may be consortia of institutions publishing data such as the eCrystals Federation, the CCDC will interact with them. In the longer term if repository Federations work, an individual or institution has to take a co-ordinating role to maintain the infrastructure. Is there a role for learned societies? Should co-ordination be mediated by committee? It was noted that some degree of permanency is offered by learned societies. There may also be an international monitoring role for IUPAC.

Most data is either: a) sent to CCDC and a deposition number issued, which is quoted in the journal paper or b) harvested from the publication source (predominantly the journal). CCDC is trying to automate the process of identifying crystal structures in papers, but it was observed that it is getting more difficult to get information from a paper because the crystallography input is getting less and less. Digital Object Identifiers (DOIs) are now routinely stored with papers.

One issue is identifying what is new in a repository i.e. trying to find out if anything in the repository has changed. CCDC may have obtained the data from another source and de-duplication is essential. CCDC see themselves as the prime source of data and aim to be comprehensive and authoritative. Structures can be acquired in several stages of the process and from numerous sources, so a versioning mechanism is needed. An example was given where one can have the same experimental datasets but different structures from two different people who give two different un-related numbers. CCDC can get a revised structure from the originator which is updated and the same number kept. This versioning information is hidden from user view. De-duplication happens if two structures are the same; more than one deposit may be associated with one structure. CCDC takes the best data and runs a provenance check via email records. So there may be a disjoint between the initial and the final structure. Once a structure is published, it gets a different 6-letter code but this can be supplemented subsequently e.g. 01, 02, 03 etc. in an open manner, once the structure has been published.

The CCDC process can be divided into two separate parts: pre-publication (working with authors) and post publication (adding value). The pre-publication database accepts 38,000 CIF files per annum. CCDC scan over 80 journals to find information and retrieve about 500 structures from Chemical Abstracts. Historically there was more searching for data, but now people send data to CCDC. About 70% of data is handled pre-publication; in twelve months CCDC estimate it will be 100%. In addition, 25% data comes from the American Chemical Society journals. Currently there is much human intervention in the data management process, some of which is adding value to the data.

One further issue for CCDC is knowing that a structure is “worth looking at” i.e. if it is a crystal structure made available through an alternative repository platform or by a different institution, there are issues of labelling and quality, of compliance with an application profile, and an issue of knowing if it is a new structure. It was observed that the landscape could “get anarchic” if structures were not fully described. However eCrystals is advocating full description. Political pressures were also mentioned: if RCUK mandates self-deposit then how will quality be assessed? Will such mandates “create more rubbish?” The user needs to know the quality criteria: data has to be fit for purpose with appropriate indicators of quality. There are quality flags in the CCDC and provenance is stated. CCDC proposed that either IUCr or CCDC could have a role in assuring quality.

CCDC noted that the proposed Federation diagram needs to be corrected: CCDC should be positioned as an aggregator, data centre and publisher[1]. It is the sole repository for crystal structure data associated with journals from some commercial publishers, such as Taylor & Francis, Wiley and Elsevier. The RSC and some other publishers do store their own data. Authors send their material to CCDC and referees approach CCDC to complete reviews in a process that is not linked to the journals. A question was asked about what other supplementary information is deposited? For example, it was noted that Taylor & Francis do not wish for pictures of spectra within the text. CCDC is not considering using a pay-per-view model but rather a subscription basis. CCDC have considered the Wikipedia model i.e. a community-driven model, but has not fully thought through the options or repercussions. They do not have RSS services, as new versions of the database are currently made available only at quarterly intervals.

Rights issues are key and the main objective for CCDC is to acquire data from repositories. It was observed that SPECTRa is using Creative Commons licences and eCrystals has a rights policy with a pointer in place. A clear identification of rights is needed, with a declaration of rights embedded in the metadata schema. It was mentioned that there are issues with METS files, which need a direct link to the CIF as in the Dublin Core standard.

It was observed that some reflection on the value of data for long term curation is essential, with an example being the images captured from a diffractometer. Not all data should be stored for the longer term. The analogy given was “like keeping stuff in your attic – when you move house you throw it away”. It was acknowledged that other areas of science may be different, such as protein structures, where data may be stored for a longer time. The CCDC has been in existence for forty years, but does not have a preservation policy. It was noted that IUCr is the first alternate store for the CCDC data, if there was a business crisis in the future. Policy has been defined by “modus operandi” i.e. they have “been doing it for last 40 years so it works”. The community also acts as a backup. It was remarked that if CCDC failed to keep the database up to date, people would “get on our case fairly quickly”.

4.2 Chemical Database Service (CDS)

Profile:

The Chemical Database Service (CDS) is based at the Daresbury Laboratory (part of STFC), and aims to provide access and search functionality to all the primary sources of crystallographic data (amongst many other forms of chemistry related data). These include the Cambridge Structures Database (CSD) (small molecule carbon containing), ICSD (Inorganic Crystal Structure Database) and CRYSMET (metals and alloys). Data are acquired through purchasing licenses to collections.

Role in Federation: supporting partner, aggregator service.

CDS have developed a prototype aggregator or harvester of eCrystals: data is harvested, indexed and made searchable alongside that from the other databases. The CrystalWeb interface is the only method for simultaneously searching all crystallographic databases. It is possible to search on aspects or components of the data (e.g. unit cell) in addition to normal ‘bibliographic’ metadata. eCrystals does not make this type of data available as part of its disseminated metadata – therefore these records must be harvested and then indexed according to CrystalWeb requirements. Software would have to be developed to index eCrystals data appropriately for CrystalWeb, and this may cause a problem in the scale up of data that would be held in the Federation. Versioning and de-duplication would also present problems.

[1] Note that this early diagram was superseded by the model in Figure 1, the Federation Model in the Dealing with Data Report and subsequent Federation schematics.

4.3 Chemistry Central

Profile:

Chemistry Central is a publisher of chemical Open Access journals and articles, with >30 sub sections and six linked journal titles in other domains.

"Chemistry Central is a relatively new service (launched August 2006) publishing peer-reviewed open access research in chemistry from BioMed Central, the leading biomedical open access publisher. The Chemistry Central Website currently features chemistry-related articles published in BioMed Central journals and independent journals utilizing BioMed Central's open access publishing services. Chemistry Central has launched the Chemistry Central Journal and is planning to launch further chemistry-specific journals".

Current handling of supplementary data: "Additional Material files should include necessary material that cannot be included in the PDF version of the published article, such as large datasets or movies. The main manuscript should include a short description of any additional files and software necessary to view them. If the manuscript is published, additional files will only be made available in exactly the same form as originally provided.

Characterization of compounds:

For known compounds used in syntheses the methods of preparation and the literature data used to confirm the material's identity should be cited. For all new compounds sufficient evidence to establish the identity and the degree of purity of the compound must be provided. Experimental data should generally be included within the Additional Material rather than within the main text of the paper and should include relevant spectral and other data. Copies of spectra used in the characterisation of compounds may be reproduced as figures in the Additional Material. X-ray crystallographic data, atomic co-ordinates, nucleic acid sequences and protein sequences should be deposited in an appropriate database in time for any relevant accession numbers to be included in the published data".

"Authors publishing with Chemistry Central retain the copyright to their work, licensing it under the Creative Commons Attribution License. This license allows articles to be freely downloaded from the Chemistry Central website, and also allows articles to be re-used and re-distributed without restriction, as long as the original work is correctly cited".

Role in Federation: supporting partner. Publisher, providing articles.

Chemistry Central Journal is an emergent journal that publishes Open Access articles (author pays) in electronic only format and as such, is keen to adopt and develop new technologies to support the process and make it more valuable and information rich.

The funding model of author pays to publish covers all costs for dissemination and adding value (peer review and additional electronic services). It is possible that this funding model can contribute, in part, to the preservation of the data.

It is important to keep/maintain a copy of the supplementary data associated with an article as:
a) there is currently not a sustainable model for funding the preservation of data held in open access, institutional, or other repositories and

b) the established methods for storing and accessing crystal structure data related to journal articles are not necessarily open access, and therefore do not fit into the approach adopted by Chemistry Central.

There was some discussion around the possibility of the journal operating a data repository for authors to deposit crystal structures related to publications in Chemistry Central. They have a vested interest in DSpace and would use that platform. They were interested in possible commercial (author pays) possibilities of publishing data, where the process of peer review of repository data is performed and a ‘stamp of validity’ issued. This might indicate conformance to the eCrystals Federation application profile.

Chemistry Central are keen to adopt the InChI when it is mature, to represent all areas of chemistry, as persistent identifiers are seen as important. The DOI is preferred for independent persistent identifiers.

4.4 International Union of Crystallography (IUCr)

Profile:

The IUCr operates partly through entities called 'commissions': there is a commission for journal publishing. There is also a committee on electronic publishing, dissemination and storage of information.

IUCr already operates as a publisher of data linked to the articles and therefore has experience and an interest in this area. IUCr is a maintainer and developer of standards in crystallography and is a potential adopter and maintainer of any standardisation process undertaken by the Federation (e.g. schemas, workflows, namespaces, terminologies). IUCr have established links with the eBank-UK Project, attending workshops and expressing an interest in ongoing work. They have participated in discussions, and finally have become an official supporting partner in Phase 3.

The IUCr has published journals since 1948. Seven titles are currently published online: *Acta Cryst. A, Acta Cryst. B, Acta Cryst. C, Acta Cryst. D, Acta Cryst. E, Acta Cryst. F, J. Appl. Cryst., J. Synchrotron Rad*. The journals provide HTML and PDF for each current article.

Metadata: No details are given in the website. The search interface for the journals supports Full text, article title, keywords, abstracts, author affiliation, author, limit by journal name and date. RSS feeds are available.

Role in Federation: IUCr is a publisher and a large and very significant source of journal articles.

IUCr consider the preservation of crystal structure data to be as important as the dissemination or publication process, and indeed IUCr journals require CIF's and structure factors (derived data) to accompany any crystal structure submitted for publication. The act of depositing data being tied with the publication process (and being mandatory), reduces the advocacy requirement. These data files are registered with a DOI as components of a scholarly article.

The IUCr journal publishing process would not necessarily be contradictory to deposit in an institutional repository and IUCr would not consider this to be contravening any journal rights. Neither would/should it conflict with other publishers processes, providing the timing of release into the public domain was appropriate. The main difficulty comes with informing and educating authors / depositors.

It is entirely possible to create eBank metadata, according to the application profile, for data contained in an IUCr publication. This could be investigated if there were community adoption of the eBank approach and it is deemed worthwhile to make IUCr publications data visible to OAI (or similar) harvesters.

The IUCr consider the role of a subject repository as absolutely crucial for (complete) preservation, and view such a facility as central to the whole scholarly process. This is not necessarily to the exclusion of the Institutional Repository; in fact it would be complementary as it is desirable to have a subject repository that contains all data in institutional repositories and other sources. This would provide a centralised preservation facility at the same time as duplicating data (LOCKSS model). There is no raw data in IUCr publications, but it is desirable for these datasets to be kept and made available "somewhere and somehow".

Institutional repositories provide a valuable testbed for interoperability, which is essential for a distributed system such as the eCrystals Federation model. Interoperability with a central archive is an essential part of this landscape. It is important to maintain a community application profile to ensure interoperability. There is a difficult case to be made for a sustainable business

model to provide and operate a subject archive funded by service provision, i.e. charging for services or funding from a public services budget. Perhaps there should be no copyright or IPR constraints on data, for it to be harvestable by a subject archive and/or any services provided on that basis.

The InChI is likely to be extremely important in digital communication of chemical structures, but is not yet mature or widely adopted. At present, it cannot cover all chemical structure types, but broadening the application is in development. Persistent identifiers are important. The DOI is preferred, as IUCr already register scholarly articles and associated derived & results data.

4.5 ReciprocalNet, Indiana University

Profile:

ReciprocalNet is a well established consortium of partners (US based but also including University of Sydney and the UK National Crystallography Service at Southampton) sharing and publishing crystallographic data by means of Open Access data repositories based at each of the twenty sites. ReciprocalNet is funded by the U.S. National Science Foundation as part of the National Science Digital Library project.

The effort is centred at the University of Indiana (Indiana University Molecular Structure Center IUMSC) and the project director is John C. Huffman. The full list of partners in ReciprocalNet is:

Indiana University, Consortium for Advanced Radiation Sources, Los Alamos National Lab, Massachusetts Institute of Technology, McMaster University, Northwestern University, Ohio State University, Princeton University, Purdue University, University of California, San Diego, University of Cincinnati, University of Iowa, University of Kansas University of Minnesota, University of Southampton, University of Sydney, University of Wisconsin, Wake Forest University, Youngstown State University.

Mission statement:

The stated remit of ReciprocalNet is to provide not only access to structures but also associated services like visualisation and also learning objects. Although at the moment they provide data that is open access only in the sense that the data is freely available (not OAI-PMH compliant), they have a stated commitment to interoperability (although it is not specified if the standard formats they intend to support are for file/chemical content or for metadata sharing.)

"ReciprocalNet will construct and deploy a distributed, open, extensible digital collection of molecular structures. Associated with the collection will be software tools for visualizing, interacting with, and rendering printable images of the contents; software for the automated conversion of local database representations into standard formats which can be globally shared; tools and components for constructing educational modules based on the collection; and examples of such modules as the beginning of a public repository for educational materials based on the collection."

Architecture and technology:

The architecture is distributed with participating sites operating common software that allows the storage of samples and metadata and the application of common services (e.g. search) across sites. A couple of specialised services act as co-ordinators e.g. to provide a network-wide search. Once a sample is identified through the search interface, the user is linked to the remote site. A site is a Web server that runs the ReciprocalNet site software and is connected to the Internet. When the ReciprocalNet package is installed on a server, the server becomes a site in the ReciprocalNet Site Network and may begin contributing to the ReciprocalNet molecular structure collection.

The site database contains metadata about samples in ReciprocalNet and is stored in a Relational Database Management System (RDBMS) like MySQL. The site repository contains actual data files for samples in ReciprocalNet and is stored on the server's file system directly.

These data files might include .CIF files, .SDT files, .ORT files, .CRT files, .PDB files, and so forth. The site database contains metadata, the site repository contains data.

Role in Federation: supporting partner.

Note: Interview synthesised from Access Grid meetings and supplemented by information collected at the CrystalGrid 2007 international workshop (Indiana, April 2007).

The ReciprocalNet (RN) Project has developed its own software platform for managing crystallographic data as it is generated in the laboratory. This has been adopted by approximately 20 sites, mainly in the United States.

The deposition process is linked to the collection and work up of a dataset and the system is primarily designed for the crystallographer to use in the laboratory. The Laboratory Information Management System (LIMS) approach is seen as the incentive or driver to deposit, and tools are provided to enable presentation of results to collaborators. RN is not designed for the dissemination of data into the public domain: it is not linked to publication processes and some see it as conflicting. However, it is possible to make data available to members of the consortium and open to the public. A subject repository is a plausible approach, but individual labs generally want to maintain their own identity and ultimate control over their data.

Access management between collaborators and research group workers within the institution is built into the software. The workflow is rigid across the RN systems, with particular files/formats required and the use of specific integrated software is necessary as part of deposit process. To some extent, a laboratory has to adopt the RN workflow if it is to use the system. Metadata for RN entry is acquired or generated during deposit process, which is integrated into the lab workflow. No metadata standard for publishing has been adopted, but in theory it is possible to align with the eBank application profile. No controlled vocabularies or keywords have been employed, as the system is not primarily designed to be a dissemination tool or support publication / linking. However, there are education and learning pages associated with some datasets, which was a condition of funding.

A related project, Common Instrument Middleware Architecture (CIMA), uses crystallography as a test-bed for remote experiment monitoring and storage of raw data. The raw data is stored on the Indiana University magnetic tape store, but no financial or preservation policies are in place after the project finishes. It is expected that this will devolve to the local institutions if the system is maintained and adopted after the project finishes. It is seen as important to preserve the raw image data. A preservation model has not been identified and therefore the long term availability of data in the consortium is not ensured and is currently considered to be the responsibility of each individual partner site.

There is the ability to control release of data into public domain and all collaborators can see a private record, but there are still problems with agreement between all parties on making public.

RN has not implemented the use of persistent identifiers. If there is a financial consideration for assignment, this would have to be included in the charge for a crystal structure determination. The US system is one where the staff crystallographer and lab are funded by charging for the service, so additional charges could be controversial with ‘customers’ and would require advcoacy of the potential benefits.

4.6 Royal Society of Chemistry (RSC)

Profile:

The Royal Society of Chemistry (RSC) http://www.rsc.org/ has headquarters in London however the Cambridge office is home to RSC Publishing.

Nature of organisation: Professional society with worldwide network of members; Publisher: "world leading in the electronic publication of Chemistry journals. The RSC publishes over 20 journals and other periodicals".

Role in Federation: Supporting partner. Publisher, providing article metadata.

Discussion focussed on aspects of publication practice. Initial exchanges examined issues around “prior publication”. It was noted that if one was reporting an exciting new structure and the crystal structure data was 90% of the paper and this data was also published in a repository, this would be considered prior publication. The RSC raised the question of “how do you draw the line between “prior publication” and acceptable practice? What proportion of an article would be contained in the Crystallographic Information File (CIF) and how much in the paper? An example was quoted in a Special Issue on Photosystem II in *PhysChemChemPhys*, **6**, 20, 2004 p 4733 Biesiakadka Jacek DOI 10.1039/b406989g. Crystal structure of cyanobacterial photosystem II. In such a case, publication in a repository would contravene novelty. This is a major issue for the RSC as a publisher.

Other approaches were discussed: the National Crystallography Service publication policy addresses four publication scenarios: accidental, traditional publication, independent data publication and independent, but linked to a journal article. New publication models such as Chemistry Central may evolve; in this case an author pays approximately £1000 to publish an article. Would a researcher be prepared to pay approximately £1000 to have a dataset reviewed and published?

One question for the RSC was raised: if you take away all data and experimental methodologies from a paper what are you left with? It was suggested that the intellectual part of a paper is the discussion. However journals such as *Crystal Engineeering Communications* are putting more and more information into the CIF and Supplementary data. How might an author incorporate up to one hundred crystal structures into a single paper? Can new approaches involving data synthesis, knowledge engineering and data mining be explored? It was noted that a paper with twenty-four crystal structures had been received for publication. How should this be handled? Is this a trend? The increased ability to mine large volumes of data may signal a concomitant increase in data submitted for publication. The publication business model may change in future to include more of the data: where would the eCrystals Federation of repositories be positioned in this landscape? The potential blurring of boundaries could create problems for some publishers. As an alternative, the RSC could work with the Federation and explore new and interesting avenues in partnership. Other stakeholders might also do interesting things to encourage data deposit and use of the original article. It was noted that the RSC is less likely to develop small niche applications than some other publishers.

There was some discussion about article size and the contraction of article size was observed with twenty-page issues appearing. It was suggested that this might be related to changing models of publishing. Rejection rates are still going up because submitted material is still increasing and papers were not innovative enough. Would this trend lead to smaller publications and with data alongside the traditional paper? Once again it was noted that the primary value currently is the intellectual interpretation and that data is property of others.

The potential of semantic mark-up in the Prospect Project[3] was explored. Other publishing channels such as people blogging their own results were mentioned, but the RSC perception is that the publication of research results currently remains largely with publishers. There was some speculative exchanges questioning whether the new generation will be more open to such tools? The RSC quoted the suggestion that they have less technical knowledge than the current generation but that they are more open to sharing materials. The notion of community sustainability was also explored: e.g. the “anarchic ontology” concept: at Hinxton (i.e. the Gene Ontology[4]), where the staff are sustaining the ontology.

The RSC is interested in approaches to the peer review of data sets. It was observed that the journal *Atmospheric Chemistry & Physics* has been using open peer review for a long time (“Public Peer Review and Interactive Public Discussion”)[5]. There was a perception that if the funding body takes a stance, then that will have a significant effect on the business model. At

the moment the RSC is looking at selling models. This led to a discussion of preservation policy. Journal back-ups are available, if the RSC "disappeared", then these would be given to the British Library. It was noted that the *Internet Journal of Chemistry* disappeared from the Web and then reappeared. The RSC view of supplementary information was then investigated. "You hold supplementary information – what's your opinion on preservation of supplementary information?" It was viewed as not core to paper but was referenced from the journal article. The RSC has made a commitment to preserve links to maintain the data, but not necessarily to migrating formats. There are various formats; most are CIFs and enhanced PDB files (for Crystallography). The RSC was asked "Do you see yourselves as having a moderator role?" It was observed that there are rules and regulations for supplementary information. They encourage people to keep material in a structured form and other publishers are also interested in this issue. In addition, the RSC were interested in adding value to the supplementary information themselves.

Publication time in RSC journals is 3-4 months. For "Rapid Communications", the publication time is shorter and depends on the time to complete the peer review process. Publication times may be extended if the data is peer reviewed or if reviewers are asked to check supplementary information. Machine-driven validation of data is envisaged or self-validation with a software tool such as OSCAR[6]. The RSC estimate that from 15-30 minutes up to one half-day per article is taken for peer review ; humans are a major block in the publication process. The RSC does not have any plans to carry out open peer review and this innovation would have to be led by the community. The prospect of open peer review of datasets was considered and the pedagogical benefits for PhD students was mentioned, however formal peer review of data may be more likely in the future. If there are clear standards for the process, then it is not viewed as a problem, however if data is published as having been peer-reviewed, then there is an overhead on that process.

International Chemical Identifiers (InChIs) are not currently published in RSC printed journals, but are being generated in the Prospect Project, and will be in RSS feeds, as will OWL representations. This was thought to be an innovation. As a first step in advocacy for the use of InChIs, the RSC will be heavily promoting this feature and Google will search/index them. The RSC is adding InChIs to current papers prior to launch of the initiative: this represents about 25% of the number possible and will be ramped up in the future. The perceived limitations of the InChI itself, is a limiting factor: they do not effectively describe coordination complexes, ions and extended structures. Retrospective allocation of InChIs will be tackled once the tools are robust and there might be individual papers and journals worth covering.

Further enhancements include the annotation of papers with terms from an ontology of analytical chemistry techniques and assays, i.e. a process-based ontology. Index terms will be collected, structures will be analysed and a candidate ontology developed for human refinement. The RSC also hope to acquire DOIs from Crossref. It was noted that eCrystals Federation should explore ways to collaborate with the Prospect Project. A direct pointer to the DOI will demonstrate linking to data at the publication stage. Such an arrangement may cause access problems at the earlier peer review stage. It was observed that we need to consider the longevity of data in repositories. Currently information can be accessed if it is directly linked but information / data can also exist elsewhere, and there is a question around who is providing the preservation service: CCDC or IUCR or some other body? There are issues associated with different types of file and varied format types, and with binary data coming off equipment such as a spectrophotometer. It was observed that in future, new algorithms may be developed which can analyse additional data types. The constraints associated with proprietary binary formats from instruments were noted.

4.7 Science & Technology Facilities Council (STFC)

STFC hosts the world's leading large scale facility for pulsed neutron and muon sources, ISIS, at the Rutherford Appleton Laboratory (RAL) near Oxford. STFC have obtained all the ISIS data from last twenty years, of which crystal data is a small but significant subset. STFC aim to capture data automatically where possible and strip metadata out of the raw files and from log

files in order to obtain the complete record in NeXus[7] file format, however there are limitations since only certain data are captured.

There are many custom-built instruments at STFC. The practice is more heterogeneous, since individual scientists write their own custom software, but there is convergence on storing structured information. However experimental processes are completely different and vary from routine experiments with the intellectual analysis largely at the end of the process to a more intellectual experimental design approach followed by routine publication of outputs. Existing methodological differences create a varying technological base including the new DIAMOND Light Source. There is a focus on providing access to the original raw data, but STFC also provides a facility for users to upload derived data back into the ISIS Metadata Catalogue (ICAT). In contrast, the TOSCA spectroscopy instrument, sends out reduced data, and raw data is never managed.

Raw data is indexed with metadata but this raises issues around standards. Defined metadata is captured at the time of the experiment for the investigator, date, time: all environmental metadata. Semantic interoperability is also an issue. There are no constraints in metadata title field. STFC recognise that dictionaries of terms are crucial but there are no name authority files. STFC do not use InChIs, however they do try to link experiments to submitted proposals, which gives additional context for an experiment.

There is a data model for ICAT, but no mechanisms for controlling what is entered into the fields and the model is based on Dublin Core. It was suggested that STFC could create mappings from ICAT to the eBank Application Profile. STFC is trying to get existing ICAT entries into a steady state. The data portal may be one way to make the underlying data public and will be a way of providing a single access point to multiple STFC repositories.

There was some discussion about ownership of data. If a person from a university does an experiment at RAL, then who owns the data? STFC is an institution and a funder, and ownership relates to the data policy. All access to ICAT will be through an API. STFC will have its own ISIS data policy with exclusivity for a three-year embargo, then the data will be “open”. However there is a need to make sure the record is complete at the point of publication. Should this requirement refer to raw data or derived data?

STFC promote “trying to make life easier for the scientist” as an advocacy message, rather than using the more altruistic global Open Access view as a selling point.

ISIS is an interesting example of use of a third-party service for a visiting scientist at a large facility. The scientist doesn’t pay for the service directly and it is free at the point of access. The scientist does the experiment, leaves STFC with structure factors, then works up the data and takes it back to their institution. The CLADDIER Project[8] Ping feature provides an option to make a link between the scientist’s data and the facility data. Bi-directional linking is required which is not simple to implement and maintain i.e. between the data and the textual interpretation. In this way, STFC believes it acts as an institution and a subject repository. It was observed that there are rights issues associated with this practice and possible conflicts with STFC data policy.

RB (Rutherford Beam) unique numbers are assigned but these are not unique to a particular dataset. No unique identifier is assigned: DOIs or handles or InChIs are not used. STFC have created the NeXus data format. Nexus files contain full descriptions of the experiment. NeXus is a self-contained file format. The neutron X-ray muon standard is basic hdf format but is still not in wide use, but has been adopted by the Diamond facility. The Target Station 2 will use this format in the new extension to ISIS, the long-established neutron and muon facility at STFC. ICAT is the ISIS metadata catalogue and provides search and retrieve functionality. For data sets, ICAT contains a header with structure factors, raw neutron counts and log files. The NeXus format represents an aggregation of the raw data file + log file + additional data and is an exchange format for experimental data, similar to the CIF. The CIF is used at the end of the experiment for the results data.

A record is kept of every experiment ever carried out at ISIS / STFC. The data is used and re-used through a data migration process across formats. This is performed within the ISIS computing group. The data has been migrated across from the VAX platform to PCs, so this

represents a real example of digital preservation. However there is no curatorial function in place i.e. adding value. STFC does not have a curation policy as such, but this topic is under discussion within the R&D context. If an experiment was rerun in ten years time, there would be new equipment with a perceived cost benefit to data curation. The refinement of data may be dependant on third-party analysis tools but there are software preservation issues associated with this practice. It is up to each facility / unit at STFC to set their curation policy. Work has also begun on an ISIS ontology and liaison with the JISC Entag Project[9] was mentioned in this context.

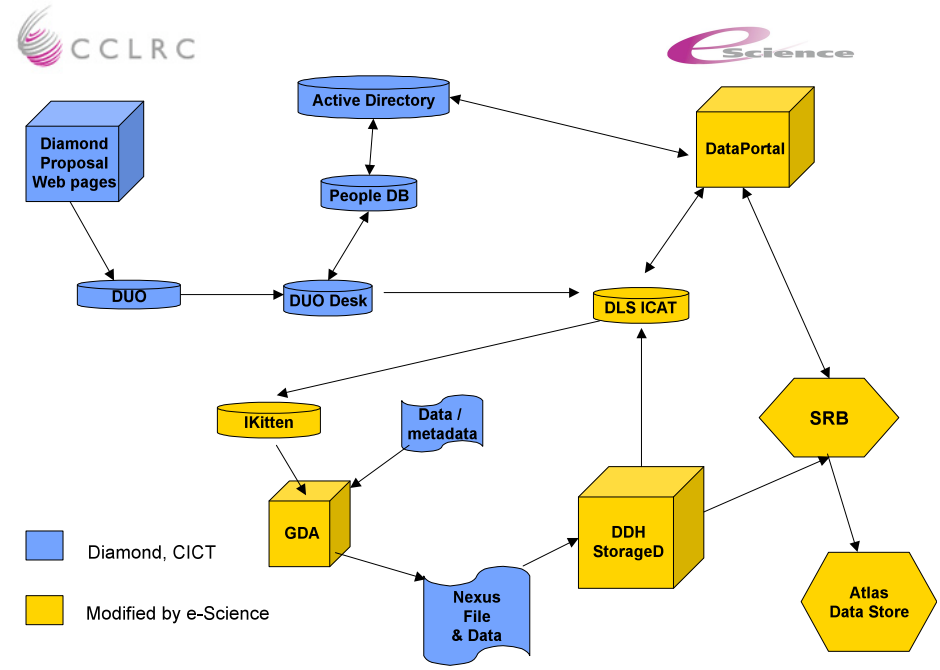
STFC wish to retain ownership of raw data in the NeXus files generated by the new Diamond facility. The user can take away the derived data and would publish the findings in a journal e.g. for a protein, the scientist would normally publish the derived data plus the model. Metadata is captured at the point of the proposal creation, and the same schema for the proposal system is used for both ISIS and Diamond: this is a result of shared development with ISIS. When the proposal is accepted, it goes into the system and into another database called “duodesk”. When beam time is scheduled, then the information gets propagated into ICAT.

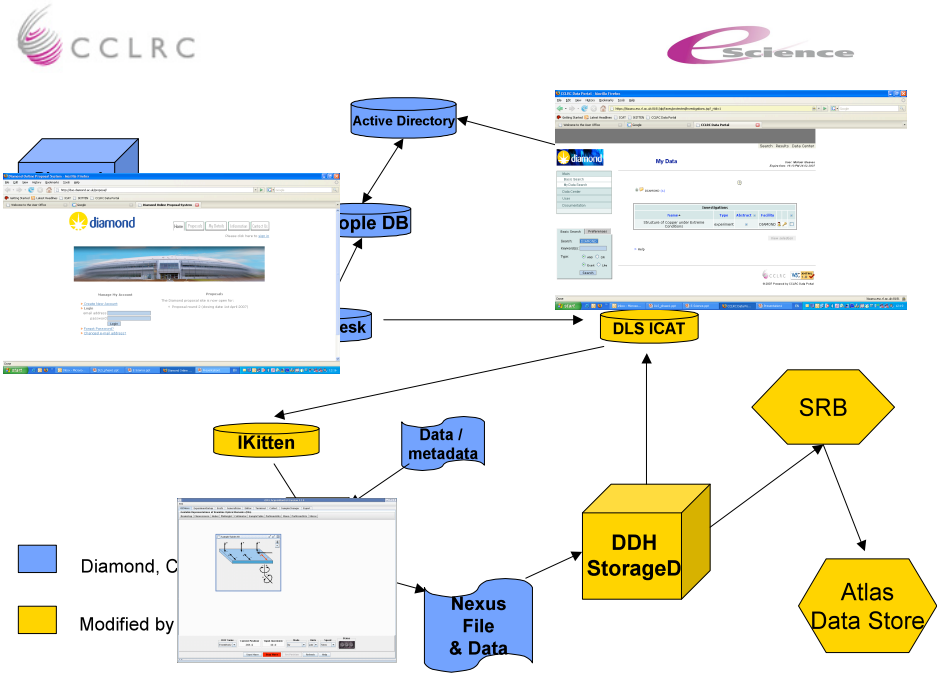
The Diamond data management system is outlined below and in Figures 3 and 4. From ICAT an instance of a record is spawned (read-only), ikitten, and is independent of the central database. Ikitten delivers record information to the beamline. A generic data tool across all beam lines (GDA General Data Acquisition tool) reads the Ikitten record and any other services, and populates the info into the NeXus file. A component in the GDA tool, reads back the NeXus file into the system and includes the proposal information and repopulates ICAT with the data. The GDA registers the file location with Storage Resource Broker (SRB). STFC plan to copy the file to the secondary data store at this point and that location is registered with SRB making a link. In the next stage, SRB puts files into containers and when they are large enough, they are sent to the Atlas Data Store (ADS). Files will be deleted off disk over time, when they are off the beamline or the intermediary data store, and SRB will only know the file exists in the ADS. ADS data storage policy currently requires payment on an annual basis. In 5-6 years, STFC estimate they will collect 10-20 TB data per day through Diamond. The data storage policy will evolve over this time and they aim to keep data available on beam-lines for a minimum of 28 days. A user will access the data from the data portal or will use SRB commands or linux commands. A user cannot delete data. This approach is subject to proof of principle and STFC is currently deploying it on an active beam-line. Images are produced as an ImgCIF crystallographic binary file (i.e. not ascii). This format is ideal for describing raw data off the machine and can describe any type of image file. Use of the Diamond ImgCIF format, is one political strategy to gain interoperability between instrument manufacturers, and there is already some buy-in. The file is metadata-rich but not all data will be contained in this file. STFC need to decide how to incorporate the data into NeXus.

STFC science departments are developing the data portal side, however STFC do not have permission to look at the data because it contains confidential information. The policy is to pay for the ADS to curate the data for the user; it remains to be seen what user demand is evident. This is the Phase 1 basic infrastructure and software. STFC need to work with the scientists to refine the NeXus format. Three synchrotron facilities have agreed on a base definition for NeXus. It was noted that the proposal must fit into the schema in the database and that some level of validation is required at this point.

After the GDA writes the NeXus file, more information is added using the command line tool being developed by STFC, and theoretically validation could happen at this point. Alternatively, the data could be validated when it is read back into the GDA. When the data is made public, it must have been validated at that point. Versioning is a difficult issue, since the data may be bespoke for the experiment and the provenance chain is very important.

Figures 3 & 4. (Slides courtesy of Alun Ashton, Diamond).





4.8 University of Cambridge / SPECTRa-T Project

The selection of DSpace as the repository platform at Cambridge was largely opportunistic, following University policy for the institutes to seek collaborative opportunities with MIT, in the mould of the Cambridge MIT Institute. The Library used DSpace as the collaborative link: Fedora didn't exist at that time. The SPECTRa-T Project[10] team have become more familiar with DSpace during the past year of the SPECTRa Project, and have had access to additional expertise. Following take-up at Cambridge, Imperial College also adopted DSpace. Cambridge

thought that eprints.org software was too limited for chemistry and not extensible. They are now investigating FEDORA. Several repository platforms are envisaged at Cambridge in future, however for now, both Imperial and Cambridge are committed to DSpace. It was noted that Imperial doesn't have its main repository up and running yet. DSpace at Cambridge is the Institutional Repository (IR) and there is a specific chemistry repository: SPECTRa which is under the control of the project. At the end of the project, it will be offered to the chemistry department. The repository acts as a holding repository, and content for long term retention will be transferred to the Institutional Repository. Data may remain in the departmental repository.

There is a similar situation at Imperial College where there is a separate SPECTRa instance. This approach has been driven by technical and project requirements. From July 2007, there will be a DSpace theses repository at Imperial. It is not yet known whether chemistry theses will go into the SPECTRa repository, or into a dedicated one. It was questioned that if a thesis has data associated with it, do you store the data and the thesis together, or separately? DSpace is not designed to connect repository instances. Establishing connections is a manual process and semantically linking objects arising from different projects is not well managed. Cambridge has a commitment to long-term preservation for its institutional repository. SPECTRa has been designed to store particular types of data for the foreseeable future i.e. for five years. Large numbers of aggregator services are envisaged in the future and concomitant pressure to develop unifying technology to link them successfully. A requirement for a central repository is not perceived unless there is a political need. There was a belief that users will not be concerned where "stuff" is stored, as long as search engines can locate it. The model for a department is for it to archive and present research outputs in a discoverable manner. There was a view that the time of a centralised disciplinary repository for data is drawing to a close, and that there will be greater reliance on institutions publishing data and associated metadata for discovery.

There is an IPR issue where the institution is exerting rights over data created by an individual. We should make data available, devise a clear licence and allow reuse without permission – like the Science Commons[11] principle. It was thought that the eCrystals Federation should not make any restrictions on data dissemination. However it was acknowledged that any policy imposed by RCUK needs to be taken into account. Long-term preservation is also important. It was suggested that if there is appropriate mirroring of data, then data will be cloned by other organisations: "If data isn't interesting, it will not be mirrored". An analogy to the transclusion principle of linking in hypertext systems in computer science was drawn. It was observed that we have to be clear about ensuring that the quality of data is consistent across all sources. If this can't be guaranteed in some way, people may be reluctant to use the data; we need quality assurance processes (QA) in place.

Access management methods for SPECTRa were discussed i.e. who can deposit data. DSpace has a concept of administrators and users in a two- level hierarchy. In SPECTRa currently, this has not been addressed beyond the DSpace tools that already exist. People surveyed in SPECTRa have asked about the ability to discriminate between different individuals who can access the data sets: supervisor, institutional staff, creator, wider public etc. This is important for depositing on a long-term basis. At Cambridge, data structures are deposited by the Crystallography Service staff. Permission is determined by the research group supervisor, and they have autonomy within the department. Cambridge will not seek to create a departmental policy on universal approaches because there is no real requirement for differential access, since deposit is carried out by a single individual. In contrast it was noted that at Southampton, multiple people are making deposits.

A universal system of validation and QA was proposed, with a minimum set of standards. Curation of the metadata is needed. This is especially important for legacy data and worth the effort to capture effectively. Two aspects were raised:
a) How good are the data results assessed by the peer review process? Is there any validity in a ratings approach? People may be sceptical about a Federation of repositories as an alternative to a central resource.
b) How well is the data marked up and is it compatible with other data elsewhere?

This point can be addressed by the metadata application profile and the degree of technical processing of data and metadata. There are issues around how to annotate a crystal file with a computation, how to control relationships, and hierarchies of metadata if people want to annotate. A protocol would be needed for the Federation: metadata is more important for a Federation of repositories. Both crystallography and computational chemistry were described as "silos" with no cross-linking. A Federation of repositories offers an opportunity to address this gap with a critical mass of data.

The process of ingest was discussed. A "golden moment" was described when the crystallographer and the chemist meet in a customer / service relationship, and the crystal data and information is handed over to chemist. At Cambridge there is a clear service provider relationship. The handover point is when the crystal is finished and put into archive as a CIF in a single upload and the main validation checks happen at this point. Cambridge have a clearly defined API for the deposition process. The Service Manager defines the protocol and business process with paper-based forms. There are mandatory files and fields. There are no electronic lab books: everything is written on paper. The files are then handed over to the chemist. It was noted that this is a very different laboratory process and approach to that followed at Southampton.

Naming structures is an issue and a limitation. Allocation of a unique identifier is much more important and the importance of the InChI was stressed. DSpace uses handles as unique identifiers. SPECTRa is diverting resources to address this problem. How is this mapped onto another instance of DSpace? The field is hard-coded into the physical machine but there is a way of registering with the handle authority to ensure persistence, however resolution is an issue.

Distinction between raw and derived data was discussed. One SPECTRa Work Package deals with NMR data. Raw Bruker files have a shelf life of twelve years and these can be dealt with through use of JCAMP file formats. FIDs are captured but there is no policy for capturing raw data. Estimates were made of 1Gb per crystal structure. This is in contrast with Southampton which stores raw data at RAL, but which is separate from the repository. IUCr are keen to get back to the raw data and there are plans to define a common format for the binary raw data.

Open repositories give third parties the opportunity to reanalyse data. It was felt that some crystallographers would be uncomfortable with this because other scientists might acquire peer-reviewed publications from re-analysing their data. There is a perceived very strong argument for capturing both raw and processed data, but it is less clear which you expose. It was noted that the basic process might be to capture both and then determine policy. Regeneration costs versus storage costs need to be considered, and the data might not be reproducible in any case. A statistic of less than 1% molecules exist and 99% of all molecules ever made don't have a physical instance as described in the "golden moment". This point is agreed in an ad hoc manner with the chemist in an individual process. It is a joint decision when a crystal is finished between the crystallography service and the chemist. Southampton has an alternative view of a laboratory data management system overseeing this workflow process. It was noted that Southampton hosts a national service and includes results from PhD students.

At Cambridge (pre-SPECTRa), the Service Manager used a manual back-up process on CD-ROM and these were collected. There was no concept of an archive or communication and open access to the data: it was a local and closed environment. This illustrates the variations in laboratory processes between crystallography departments and services, and is a key issue for the Federation. In contrast, the High Performance Computing Unit, has recognised this type of research data as a source which needs archiving and curation, and this data is now captured at source and transferred to an archive at the "press of a button". Others in computer science ("command line people") were perceived to be at the other extreme, whilst synthetic chemists and spectroscopists were perceived to be "in between" with their file store approach. Such arguments for a repository Federation makes it easier to drive home change in an institution.

SPECTRa is using the eBank schema but in a slightly extended form with annotations for computational chemistry and spectroscopy i.e. using the eBank application profile as a core profile. To achieve uniformity in the presentation of the record, SPECTRa use METS packages,

where the record looks the same but different metadata is exposed, such as the formula. However, storing big binary files is seen as a problem for DSpace.

On the topic of terminologies, it was thought that a controlled vocabulary is useful for precise information, but there is now a perception that free-text indexing is just as powerful and in any case, “users hate putting in keywords”. There is currently a mixed model of formal and informal terms. Computational chemists have different perspective and are using free text and set terms. It was noted that this is a long way from ontologies. SPECTRa advocated use of Knowledge Organisation Systems (KOS) and there are taxonomies for computational chemistry, which will be part of SPECTRa. This is seen as providing a flexible and quasi-democratic way of managing terminologies.

Finally, there was a discussion about institutional mandates. On July 1st 2007, Imperial College (IC) de-federated from the University of London, and students choose which degree they will be awarded. If they choose an IC degree, they will be mandated to deposit their PhD thesis in a repository. It was observed that a strategy is needed to assimilate SPECTRa-T (Theses) and associated data over the next two years; it was felt to be too early to formulate this now. They plan to embed deposition into institutional requirements. IC is also hoping to implement a mandate that published papers are to be deposited in an IR. There would be an opportunity to mine the thesis and make a link to the repository data. In future, an automated process will be needed for preprints in order to give context to data structures. The SPECTRa team realised that the design of DSpace didn’t provide for complex objects which need to be semantically linked. This is one of the limitations of DSpace. A prediction that in five years time, all journal articles will be self-describing so that they can go into any archive, was presented. At Cambridge, there have been discussions with the Board of Graduate Studies and a mandate is estimated to be probably 2-3 years away. It was noted that there are no other deposit mandates happening at the university level; however mandates may be implemented at the departmental level.

4.9 University of Sydney, Australia

Profile:

The School of Chemistry is the lead partner in the "Molecular and Materials Structure Network", funded by the Australian Research Council http://mmsn.net.au/. The goals of the project are to link instrument facilities, data repositories and collaboratories via the grid in a research network and collaboratory.

“The establishment of these network services and the 21st century collaboratory they will constitute, will significantly enhance research endeavours in chemistry, materials science, biology and computer science, and will catalyse the formation of new linkages between these sciences. Input from the user community represented by the diverse membership of the MMSN will ensure the collaboratory has real word functionality and is user friendly."

The primary goal of the Network is the building of a new and powerful e-Science tool that will ensure that Australian scientists are exceptionally well equipped to push the global leading edge of any research that depends on a knowledge of structure at a molecular level.

The MMSN will collaboratively develop two closely related internet network services to foster and advance molecular and materials structure e-Science and its diverse application and utilisation in the broader scientific community."

"The MMSN will develop the world's first Grid-based collaborative molecular visualisation system, such that multiple users in different locations can simultaneously interact with a synchronised molecular display. The system will provide geometrical analysis capabilities, and multiple rendering options would be provided. Users would communicate visually and verbally through the network, or with conventional telephone or video conferencing equipment (http://www.polycom.com/home/)."

"Additionally there will be a database network with exceptional service capabilities that will prove invaluable for the structure sciences. The database network will incorporate and extend new developments in visualisation and analysis, and will be suitable for access from a National

Digital Library. The remote instrument access and database networks will be Grid enabled to leverage the benefits of the Grid, and to provide linkage into the emerging global Grid."

"A structure database with cross disciplinary content and powerful visualisation and analysis capabilities will exemplify "smart information use". Encompassing physics, computer science, chemistry and biochemistry, and catalysing interaction across these disciplines, the MMSN will impact all four National Research Priority 3 goals, and will be linked to other national and international Grids to become part of the emerging global Grid."

"The intent of the program is to incrementally establish a national molecular and materials structure database service, primarily serving the community represented by the MMSN. In addition to conventional databases the program will explore a Grid-based structure database system, and a complementary Grid based spectroscopic database. The database service will be collaborative in character, with the principal vehicle for collaborative interaction being the world's first Grid-based, mutually interactive molecular visualisation and analysis system. The system will provide synchronised displays to multiple monitors that may be located anywhere in the country, or overseas. The MMSN database service will be piloted by providing access to the principal databases used by the molecular and materials structure sciences; the Cambridge Structure Database (CSD), the Inorganic Crystal Structure Database (ICSD), the Protein Data Bank (PDB), the Metals Data File (MDF) and the Crystal Data Identification File (CDIF). The national molecular and materials structure database service will be established in close collaboration with the UK's Chemical Database Service (CDS), which provides a comprehensive structure and properties database service at no cost to subscribing academic institutions (http://cds.dl.ac.uk). The Australian database service will be modelled on that provided by the CDS.

Role in Federation: supporting partner, institutional data repository, modelled after eBank/eCrystals repository.

The University of Sydney crystallography department is a service facility. The Director of the Crystal Structure Analysis Facility (CSAF) is also manager of the Molecular and Materials Structure Network (MMSN), which is concerned with remote experiment steering and control at Central Facilities Labs (new Synchrotron and Neutron sources), and the management and storage of the data arising. Preservation and persistent availability of raw image data is most important for this project and the management of derived and results data is considered to be the responsibility of the home institution/experimenter.

MMSN and CSAF have not been considering the institutional data repository model, but are a member of ReciprocalNet but are keen to adopt eBank software and become part of the eCrystals Federation. As part of groundwork for this work, the issue of sustainability and preservation has been taken up with University of Sydney Library, who are keen to participate and contribute to eCrystals Federation project. CSAF considers the dissemination and integration with publishing aspects of institutional data repositories to be especially important and is keen that any software / system implemented should not impinge on the established working practices in the laboratory.

5 Synthesis and Discussion

The Synthesis is presented in twelve sections: Institutional Repositories Policy and Practice, Crystallography Laboratory Practice and Workflows, Technical Interoperability and Standards, Metadata Schema and Application Profiles, Semantic Interoperability, Data Citation, Identifiers and Linking, Federation Architecture and Third Party Services, Rights and Licensing, Data Quality and Validation, Preservation, Curation and Sustainability, Community and Inter-disciplinary Interactions, Collective Intelligence and Open Science.

5.1 Institutional Repositories Policy and Practice

Whilst there is a growing body of work relating to institutional policy associated with document repositories, there is as yet, little evidence that institutions are examining the curation and

preservation of primary data within their Faculties, Schools and Departments. The *Dealing with Data Report*[12] (2007) recommended that "each higher education institution should implement an institutional Data Management, Preservation and Sharing Policy", and the findings of this study re-enforce this assertion.

Institutions vary in their size, character and structure, and a one-size-fits-all data policy model is unlikely to work in this context. The *RIN Data Stewardship Principles*[13] provide an appropriate framework into which institutional data policies can be positioned, however data policies need to be developed locally and reflect organisational requirements and repository maturity. In addition, the data policy should reflect the repository model in place within the organisation (i.e. institutional and/or departmental).

Within any scientific research organisation, the granularity of operational structural units need to be considered (we will return to this later) and the policy-setting process should be based on critical elements of consultation, transparency, flexibility and review. Ideally policy-setting should involve key stakeholders within the community and in the case of primary research data, the inclusion of practising researchers as Faculty representatives in addition to repository managers, is essential. Existing commercial agreements may however, place constraints on the degree of flexibility in institutional data repository policy-setting: these agreements may be related to the software platform, development partnerships or data re-use. The broader range of stakeholders will also have a view; their engagement at an early stage is vital to ensure buy-in and to assist with the advocacy and dissemination process across the whole organisation.

Since this is a fast-evolving field with high-profile policies emerging e.g. NIH[14] and Wellcome[15], there is a need to review policy on a regular basis. The example in this study of the de-Federation of Imperial College from the University of London, provides good evidence for this changing landscape. Repository policies including those that embrace data, must accommodate any wider institutional mandate. The Imperial College mandate related to deposit of PhD theses, demonstrates that data repository policies must be flexible and dynamic to accommodate such supportive statutes from the highest levels within an institution. In such situations, timing of project development implementations is critical and must fit with emergent institutional policy. In some cases, mandates may be easier to implement at the local / departmental level. Finally, for any policy to be effective, there needs to be a level of compliance with policy statements and technical requirements: there may be a need to oversee these parameters and ensure that they are embedded in routine day-to-day research practice.

In the case of the University of Southampton as the publisher of the records in the eCrystals Repository, a University Preservation Working Group has been established which is carrying out a survey of the quantity and diversity of data generated by all laboratories and research groups across the University. This is in line with the *Dealing with Data Report* Recommendation for "JISC to develop a Data Audit Framework to enable all Universities and colleges to carry out an audit of departmental data collections, awareness, policies and practice for data curation and preservation". The JISC has subsequently awarded funding for the Data Audit Framework development to the University of Glasgow, as a partner in the Digital Curation Centre.

Institutional policy must incorporate access management for data repository content. Questions are often asked by academics about "one-stop-shop access," however the drawback with this scenario is data ownership. Notionally the data deposited is "open," but it is actually stored in a repository behind a firewall, and there is an important distinction between ownership and openness. Data from the University of Southampton is exposed to third party services, but obtaining permission to make the data public in this way is a key step, and not necessarily implied by simply being available on the Web. Winning the hearts and minds of researchers is essential and advocacy has a key role to play in achieving this goal. Managing data ownership effectively is critical: the time when academics are less likely to give permission for data publication, is within the first year, when patents and journal articles are pending. Additionally, the academic may move to a new institution, which may also cause a block on data publication and raises further access management issues around legacy data.

The varied nature of laboratory practice within the same discipline is discussed in Section 5.2, and this will largely determine local data policy at the research group / departmental level. Institutional data policies must be flexible enough to support a range of scholarly and research

working practices whether data creation and collection is laboratory-based, field-based, simulation-driven, from observational collection or performance-based. In crystallography, there are also relationships between datasets in a laboratory repository, data in the subject repository, related data held at remote large facilities such as Diamond or ISIS and with other third party services and data outsourcing arrangements e.g. with the ATLAS datastore. Institutional policy should reflect this complex landscape and account for the management of raw and processed data, data versioning, linking, IPR, embargo periods, data duplication and provenance. The inter-relationships between the diverse policies developed by the various players in the landscape need to be clearly articulated to assist the creator and user of a dataset, whether for human and/or machine interpretation and translation.

Policies should encapsulate all stages of the curation lifecycle . A rule-based approach may be advantageous in managing cross-organisation issues. The iRODS or iRule Oriented Data System[16] initiative led by SDSC, provides a rule-based policy development framework and "adaptive middleware" to assist with the large-scale management and preservation of datasets. iRODS is being implemented in the UK by the JISC Architecture for a Shibboleth-Protected iRODS System (ASPiS) Project[17] .

5.2 Crystallography Laboratory Practice and Workflows

In general within crystallography, there are two types of experiment: a) the 'routine' experiment where some intellect is required in the later analysis stage; b) Intellect is required in the initial design of the experiment, but the subsequent publication of results is relatively routine. Whilst within this broad division, there may be many different experiment types, there is an emerging need for a common platform for storing the results, despite differing technology bases.

Additionally in crystallography (and also in many other disciplines), there is a great difference between raw, derived and results data. This diversity greatly affects the method of archiving and the approach to dissemination. It is becoming increasingly clear within the community, that there is a requirement to store much of this data for effective preservation and curation, and full realisation of experimental provenance. Raw data presents a storage challenge and is relatively large in size, frequently in proprietary binary format and in many circumstances, could potentially be stored for only a short period of time after a validated result has been successfully completed. Conversely, derived data, whilst relatively modest in size, varies hugely depending on working practice, but the final result is invariably just a few kilobytes in size and in a community-wide standard format. Different approaches to archiving are apparent even within a single discipline. In the SPECTRa Project in areas of analytical chemistry, all the data is captured and then policies are considered on a case-by-case level.

It is important to note that there are very different approaches to working practices between larger-scale production-level crystallography services and smaller departmental installations. These differences arise from the need for a more structured approach when providing a service used by many scientists, as opposed to the procedural flexibility possible when a single researcher is operating his/her own laboratory. Larger, centralised organisations have generally considered archiving and data management, however departmental services frequently don't have any archival or access policies in place. Additionally the departmental-level service is further complicated by the issue of ownership of the data, due to the 'customer-service provider' relationship that often exists between the crystallographer and the provider of the crystal sample. This relationship can range from being an open collaboration, to a situation where the sample provider pays for the service, and the data (and ownership) is handed over to the 'client'. Many departmental-level laboratories do not have the time or resources to implement data management systems and hence are still very paper-based and operate in a 'manual' mode. There is great sensitivity regarding the scientist provider – crystallographer relationship, and data release to the public / data publication is often determined by the scientist provider. Hence there is a delicate political issue regarding data openness and accessibility which needs to be clearly defined from the outset of the experiment, as it affects the eventual availability of the results.

Whilst there are automated systems already in existence, they are generally designated "Laboratory Information Management Systems[18]" or LIMS, and publication or dissemination is

not specifically considered as part of their systems design. Any descriptive metadata collected will be poor as a result of this omission, however a LIMS system such as that operated by ReciprocalNet, provides the ideal mechanism for the capture of this metadata at the point of generation. There are advantages and disadvantages in adopting a LIMS approach in the laboratory. The implications of adopting a LIMS approach are that it prescribes an inflexible workflow that necessarily must be adhered to: many researchers find this restrictive and must alter their working practice to conform with the system. Conversely, whilst LIMS provide valuable automated processes which enhance the productivity of routine laboratory procedures, existing laboratory management systems need to be better integrated into the workflow of the experimental process they support. A prime reason for the lack of broad uptake of LIMS systems is that they are often pieces of software that reside on desktop computers and therefore do not feature within the laboratory environment. There are a large number of commercial LIMS solutions on the market (see http://limsource.com/products/products.html) with a recent focus on Web-based and distributed functionality, however generally these still fail to fully integrate with the laboratory equipment and practice. This lack of integration causes even more resentment from the laboratory researcher as a 'normal' laboratory notebook must still be kept and this would then have to be written up in the LIMS system afterwards. Thus not only does a LIMS system constrain the workflow and practice of the researcher in the digital environment, but it also fails to be of any practical interactive use in the laboratory.

Modern centralised facilities and services are now supporting such systems, however these are bespoke integrations suitable only for service providers with resources to support the infrastructure required. A number of manufacturers of high throughput analytical instruments now provide a degree of LIMS support in their software, however these systems are very much proprietary and do not offer support across a range of analysis types or manufacturer brands, and still offer only limited integration with physical operations in the laboratory. Currently there is little or no support of this type in the field of crystallography. There are also a number of LIMS that take a project view, as opposed to an instrument or technique view, and whilst they provide a record of experiments, information and data relating to a compound or project they are still proprietary and somewhat divorced from the laboratory. To provide a rounded solution that encompasses both approaches would require considerable expert integration and development work on a departmental level and due to logistical, management and financial issues, such systems do not exist in the academic sector. Further work on LIMS is recommended in Section 5.3.

For centralised research data services within a School, Faculty or department, best practice in metadata capture must begin at the stage where the research grant proposal for funding is produced. Researchers should be made aware of the facilities that they will require for information capture and preservation at the outset of writing a grant proposal and how they might apply for this support in the proposal or via other supporting means. It is then vital to continue this ethic through the setup, operation and completion phases of the project. For the departmental crystallography service, metadata capture starts at the individual crystal sample submission stage. It is clear that shared tools for the provision and addition of metadata relating to both the prior synthesis process and the crystallographic experiment are necessary and should be provided to the author or sample originator at the outset, and deployed throughout the experimental process. The crystallographic experiment is both a practical one, where data is collected on a sample and an in-silico one, where the data is worked up into a crystal structure. There is a massive loss of information relating to the experiment in both aspects of this work. In the laboratory, little or no information relating to sample manipulation, environmental conditions, instrument operating parameters or data massaging is captured, stored or disseminated. Additionally the process of working up the data into a crystal structure is an iterative one and no information is saved from each individual step to indicate the process that has been undergone. These shortfalls in recording information can often have very serious implications for the interpretation of a result, its validity or the ability to reproduce the experiment.

The embedding of repositories within the research workflow is of critical importance to their success. The development of Virtual Research Environments to underpin the research process may provide the necessary platforms to co-ordinate and streamline the various elements of the research cycle, however it is crucial that these are underpinned by a solid infrastructure for the

capture and storage of experimental data. One example is the Research Information Centre[19] (RIC) being developed by the British Library in partnership with the Technical Computing Group at Microsoft. Whilst this is currently focussed on the biomedical researcher, one can envisage a similar set of tools being provided as a chemist's workbench, and which would include tools for the management of datasets generated during experiments in the laboratory, such as the R4L[20] repository. There is also a clear need for a well-defined data deposit API for disciplinary/sub-disciplinary or departmental/institutional data repositories.

Recommendation 1: JISC should provide guidance to support the development, interoperability and sustainability of sub-institutional repositories, such as those at departmental, research group and laboratory levels.

5.3 Technical Interoperability and Standards

The crystallography community has developed and uniformly adopted the CIF format for a crystallography results data file, which describes the crystal structure result in terms of atomic coordinates and geometry, crystallographic and chemical parameters, experiment parameters and software employed. Furthermore, this format is the basis of the CheckCIF validation mechanism, which automatically assesses the CIF for correctness, internal consistency and chemical sensibility. Additionally, courtesy of CCDC, there are freely available cross-platform tools, enCIFer and MERCURY[21], for manipulating, viewing and interrogating CIF files. The CIF format has also formed the basis for innovative authoring of a machine readable publication: the format is extensible in that it enables insertion of text sections of mark-up to be added to the file. The freely available software PublCIF[22] provides a dual view of the CIF where the 'raw' mark up file is displayed in one pane and the rendered 'word processor style', with associated formatting etc tools, is displayed in the other and either form is editable, with automatic updates.

The crystallography domain is a good exemplar of "small science" best practice within a community where there is a relatively limited range of experimental types and a high degree of agreement and organisation. As a result, adoption of the CIF has enabled a degree of interoperability and data sharing between many different stakeholders, including researchers and publishers alike. A similarly rigorous approach needs to be adopted for describing the raw and processed data. There also needs to be robust metadata for preservation and dissemination. The capture of additional metadata to enrich descriptions of datasets, should provide a sound basis for similar small science operations.

In a further innovation many large scale synchrotron and neutron facilities around the world have adopted the NeXus Standard which defines the experimental framework for this type of large scale facility, providing a method to encapsulate and describe the experiment. These systems and formats support the central facilities experiment from the proposal stage all the way through to the provision of results i.e. CIF + proposal + instrument descriptions. The NeXus standard supports collaborative science carried out at large-scale facilities such as those at STFC, where visiting scientists book a scheduled time slot to use the shared instrumentation.

Within the crystallography laboratory, there are numerous different proprietary (binary) formats used by instrument manufacturers for raw image data, which can only be read by the experiment control software. There are 5-6 principal manufacturers that supply a range of off-the-shelf diffraction instrumentation, each company (and often diffractometer type) with a different raw data file format. However, particularly at central facilities where innovative developments are fostered, there are a considerable number of 'home grown' formats. Whilst the standard ImgCIF[23] has been in existence for ten years, it has not achieved wide adoption due to a particular laboratory owning and using instruments from a particular manufacturer, and instrument software not supporting the format. This manufacturer "lock-in" creates real barriers to data sharing and more work is needed to investigate issues in this area, to provide community advocacy and to promote community standards such as ImgCIF. However, a number of prominent organisations and initiatives are now implementing LIMS or experiment monitoring / control systems based on the ImgCIF standard (Diamond[24]; the Australian synchrotron[25]; CIMA[26]) and the academic community is now beginning to consider the publication of raw diffraction images in an interchangeable format[27]; TARDIS Project[28], CrystalGrid[29]).

As described in Section 5.2, the crystallographic experiment is generally split into two parts: the collection of data in the laboratory and the workup of data in a personal computing environment into a structure. This causes a disconnect between the PC and the laboratory components. Modern instruments implement and record details of prescribed workflows to collect and correct raw data, however this is considered and managed separately from the workup of the data into a structure. Workup is currently a poorly managed process, with the storage and management of files being determined by the user. The refinement of a crystal structure in the workup stage is an iterative process where subtle changes to the model are made and successively tested – generally this process is not captured, as there is currently only interest in analysing or publishing the final result. There are a handful of software packages for the workup of a crystal structure, all of which are developed by academics and generally unfunded. Accordingly few of these software packages are kept up to date or formally supported. The chemical crystallography community, (through tradition more than any other reason), widely (80-90%) uses the SHELX software package[30], which has become a de-facto standard. The SHELX package was first released in 1976 and continues to be developed by its sole author, Prof G.M. Sheldrick. The main competitor to this package, CRYSTALS[31], has received slightly more attention in the past decade, however this too was developed in the 1970's/80's and its future is in doubt. Development of this software is impractical to support in this mode and there is little sign of a new generation of crystallographers willing to write code in their spare time for relatively little reward or recognition. Manufacturers have shown a degree of interest in workup software, but this has not been widely adopted due to the open source nature of the code and the inflexibility of its deployment. The community has recognised this problem and the OLEX Project[32] is taking the code-base from these software packages and 'future-proofing' it. This initiative is very promising in that there is an attempt to record all outputs from the workflow and preliminary discussions regarding its compliance with data repository protocols have been conducted.

Similarly there have been initial discussions with publishers (IUCr, RSC, Chemistry Central, Nature), regarding the incorporation of data repositories into the publishing workflow. This is an innovation that would provide a serious driver for researchers to adopt repository methodologies and would truly be a large step forward in the accuracy and reproducibility of the reported scientific data in the literature. It is also vital that research institutions are engaged in this process, as they will be responsible for the long term preservation of this data. The eCrystals Federation is engaging libraries and information services in this area and JISC has funded a number of studies aimed at informing and assisting institutions in making financial and policy decisions relating to data repositories.

Recommendation 2: JISC should consider funding an investigation of "laboratory informatics" including LIMS, to identify opportunities for more generic workflow integration and pervasive systems to capture laboratory data and metadata in-situ.

5.4 Metadata Schema and Application Profiles

There is clearly a requirement to capture the fullest details of an experiment and to be flexible enough to account for unusual or bespoke experiments. The eBank Application Profile (AP) was developed over the course of the three phases of the project. It is based on Dublin Core and acts as a generic set of discovery metadata. It was specialised to allow crystallography-specific terms to be incorporated into the vocabulary to support crystallographic-specific services e.g. InChI searching. The eBank Project metadata schema is documented[33], and is comprised of XML schemas that allow automated validation of compliant schemas and human readable descriptions. We have treated the eBank metadata Application Profile as a "core profile", which can be built on and extended for local and/or personal needs.

In the SPECTRa repository, metadata schemas are based on the extended Dublin Core schema published by eBank for the eCrystals repository. Limited local extensions have been adopted, for example to distinguish between the originating chemist data owner ('*creator*') and the spectroscopist/crystallographer ('*contributor*'). Embargo information is also encoded. As much metadata as possible is created automatically when the deposited files for each of the three chemistry areas studied, are read by the appropriate validation processes. As some fields

are defined as mandatory, such as *embargo period, author names*, these are additionally prompted for by an "AddMetadata" page in the deposition process if they are missing: deposition cannot proceed until these fields are filled manually. This experience suggests that whilst the eBank AP may function as a core profile across the partners, the Federation must have flexible metadata policies to enable local extensions to be implemented, to successfully manage local laboratory practice.

A further example where flexibility may be required within any Federation of repositories is the requirement for packaging data *i.e*. to associate a number of data files together with some technical and descriptive metadata. This was considered during the SPECTRa Project. At the time, the main alternatives were to use RDF, METS or MPEG21/DIDL (the last two both being XML-encoded approaches). Of the three, METS was chosen because it was the simplest technology that met the requirements, DSpace supports a METS profile as its primary package format and it had already been adopted by the eBank Project.

The emergent OAI-ORE[34] initiative may provide a more appropriate model for describing compound digital objects, such as these crystal datasets. Alternatively the Scientific Compound Object Publishing & Editing System SCOPE[35] also provides a framework for describing complex scientific objects based on OAI-ORE principles. This aspect of the Federation's operation is being explored as part of the Microsoft Research funded eChemistry[36] Project.

The development of metadata application profiles for different types of repository content poses a question about the management of such schemas. The eBank/eCrystals profile could be managed within the Information Environment Metadata Schemas Registry (IEMSR)[37] at UKOLN, or within the wider disciplinary community. How does the Federation ensure compliance? It was observed during the interviews that this is embodied in software development, becoming a *de facto* standard through adoption. Whilst there are technical interoperability issues to solve through technical solutions such as applying Dublin Core (DC) application profiles, the main barriers to deposit are perceived to be socio-economic ones, raising questions such as "who owns the data?" It was also observed that we need to maximise the visibility of published data: and no-one knows how much is <u>not</u> published, although there are anecdotal estimates of the numbers of molecules synthesised but not described. Metadata describing the crystal structures is exposed for harvesting by a range of aggregator services including the eBank aggregator service and CrystalEye[38]. We can envisage a growing range of third party aggregation services based on RSS or ATOM protocols crawling the content repositories and presenting the results to the (human or machine) user, and this is discussed further in Section 5.7.

5.5 Semantic Interoperability

A Report of an analysis of the semantic issues associated with eBank UK has been published[39]. There was considerable interest across Federation partners in the potential of appropriate ontologies and the use of controlled vocabularies to enhance the value added to an article or dataset, and to give additional context to enhance the functionality of crystal structure aggregator services such as those of CCDC. The Royal Society of Chemistry Project Prospect is developing an ontology to semantically enhance textual content in published papers and is considering the development of a process-based ontology. The RSC were supportive of further collaboration with the Federation. A different approach has been successfully adopted in bio-informatics, where a **c**ommunity-maintained ontology (the Gene Ontology[40]) has found acceptance as a model. A less formal approach is to use social tagging to describe datasets for global discovery. The JISC-funded Enhanced Tagging for Discovery (Entag) Project[41] is exploring the relationships between formal structured Knowledge Organisation Systems (KOS) and less formal methodologies.

During the interviews, it was asserted that "users hate assigning keywords and don't know how to do it". Whilst this may be the case, there is a clear requirement for assistive mechanisms to facilitate the description of datasets to enable discovery. Currently there are no established name authority files to support author fields, and disambiguation when cross searching is a problem, since different terms have different meanings in different fields. The use of text mining tools such as those from the National Centre for Text Mining (NaCTeM)[42], to assign, analyse

and disambiguate terms is an increasingly attractive alternative approach and more work is needed in this area.

Recommendation 3: JISC should support further work to explore alternative and/or automatic assignment of terms and keywords to data sets for enhanced discovery.

5.6 Data Citation, Identifiers and Linking

The *Dealing with Data Report* identified the importance of "robust, bi-directional interdisciplinary links between data objects and derived resources." This included the ability to link between related datasets and between (supplementary) data and the derived journal article. Other examples of useful links include from a dataset to the funding proposal, project plan, experimental protocols and methodologies, results (raw, processed, derived), images and 3D representations of structures and textual interpretations of the outcomes, which might be presented as blog posts, wiki pages, pre-prints, reports and/or formally-published peer-reviewed articles. Such comprehensive bidirectional linking is hugely important in achieving the value chains that underlie the scholarly Web, but is very difficult to maintain, since there is a need to support cross-linking across disciplines e.g. chemistry and biology, and across sub-disciplines, such as crystallography and computational chemistry. Scaling-up is a major problem as demonstrated by the highly complex crystallography landscape with multiple aggregators and multiple data sources.

How should we assign a persistent identifier to a crystal structure? Can assignment be automated? The eBank-UK Project explored the use of Digital Object Identifiers (DOIs) for the data structures generated in the laboratory. International Chemical Identifiers[43] (InChI) were also assigned to each structure. The InChI adds value, as it provides essential chemical context, and is likely to underpin automated search and analysis, however the InChI is still under development and doesn't yet cover all areas of chemistry. The DOI provides an important link into the publishers' mode of operation, but there is a cost associated with assignment, which must be incorporated into any business model for partners in the Federation. The cost needs to be included in the 'charge' to synthesise and analyse a crystal structure. Other partners use different identifiers e.g. CCDC assigns a deposition number on receipt of the data file, but the structure is then given an additional identifier in the public database. Such diverse practice raises questions of interoperability across the Federation, and user policies need to address these issues, to promote best practice to enable discovery and reuse of the crystal structures. In addition, process metadata can contain identifiers, but these aren't necessarily unique to the results and therefore of limited use. They may be facility specific, e.g. Rutherford Beam numbers at ISIS/STFC. We begin to see the prospect of a hierarchy of multiple identifiers emerging, with local, domain-based and global identifiers assigned to a single dataset.

Versioning of crystal structure data presents a special problem to a repository and to the Federation more widely. Data describing a crystal structure may be replicated in a repeat experiment i.e. duplication. Alternatively, data may be transformed or re-analysed to create a derived dataset which is related to the original data object. The application of time-stamping and allocation of an identifier, is critical, designating the primary source of the original data from which other datasets or models have derived. This helps to establish the "first to invent" imprimatur, can facilitate embargo practice, can give an indication of intellectual ownership and authority, and enables the provenance of the dataset to be demonstrated and subsequently tracked. However, more than just a time-stamp is required. The peer review process can lead to subsequent revisions of a crystal structure and re-submission to the same journal or submission to different journals with different requirements, which in turn can cause new identifiers to be issued.

Chemistry has a particular identification problem associated with chemical nomenclature, which is difficult to assign and highly complex. Whilst the IUPAC supports a set of nomenclature rules, practical interpretation can vary from chemist to chemist. There was a view that assignment of particular chemical names should not be mandatory within the repository metadata, as it may deter the depositor from adding their data to the repository.

5.7 Federation Architectures and Third Party Services

5.7.1 Levels of Service

The eCrystals Federation concept arises from a vision of an online environment that facilitates the seamless exchange and discovery of information resources, related to the discipline of crystallography. It is hypothesised that a collaborative approach is required to improve access to resources such as data or publications from crystallographic determinations. The approach may benefit from agreements on transactions, shared infrastructure, and policies, in order to maximise the discovery and use of resources. In practical terms, the eCrystals Federation will be made up of different sources of data or literature, and services, owned and managed by a variety of organisations (e.g. publishers, academic institutions). The participants in the Federation consist of a network of loosely associated players, related to each other through interest in, ownership of, or management of resources or services relevant to crystallography. Furthermore, interactions with other organisations or services operating independently outside of the Federation may need to be considered in order to meet the full expectations of end users. This section of the Report addresses the implications for the shared technical infrastructure, consisting principally of resource storage, information exchange and discovery systems. The current technical operations within the members signed up to the Federation are considered; some design choices and protocols for exchanging information and data are then reviewed.

The approach taken by the eBank-UK Project in improving the accessibility of crystallography-related resources can be considered to be two-pronged. On the one hand, the project has described and implemented infrastructure that supports the management of crystallography datasets at the point of creation. The infrastructure is based on repository software which enhances the management of data by (1) providing a central point for data deposit, tailored to crystallography data, which can be integrated into the workflow, so that the datasets can be ingested into a managed system, instead of remaining isolated within individual collections (2) collecting information about the datasets (metadata) in a systematic manner and (3) providing a browsable archive of the datasets which grants access and download to the human user from one central point.

The second aspect of curation of crystallography data demonstrated by the eBank-UK Project is the integration of the repository into a wider infrastructure, one adopted mainly by digital library communities, which is geared towards making the repository resources more easily discoverable by third party services. The approach centres on the generation and subsequent sharing of metadata. The metadata acts as an advertisement of the existence of the data, and provides paths of discovery into the data when exposed in alternative locations and services. The metadata can be interrogated by end-users to assess which datasets are available for further exploration.

Whereas the first achievement relates to better local management of data, the focus of the second achievement is interoperability, that is, viewing the local collection as part of a bigger whole. It is the belief that the whole is greater than the sum of the parts that underpins the Federation approach. The value of the local collection is seen in the context of other related resources; within this larger collection, previously unknown connections can be tested and discovered – the local collection must not be viewed in isolation, but as part of a network of inter-related data stores. It is this Digital Curation Centre perspective on curation as described below, that the Federation must consider:

"*Digital Curation itself is the active management of data over the life-cycle of scholarly and scientific interest; it is the key to reproducibility and re-use. Metadata for resource discovery and retrieval are important, with mark-up on time/place referencing as well as subject description and linkage to discipline based ontologies providing key research foci*."

The focus is on discovery services, which are points of interaction with end-users, where information needs are expressed through queries, datasets of potential interest are identified, and then further accessed and explored if the information need is likely to be met. Underlying the discovery mechanisms is a technical infrastructure that can support varying degrees of information exchange, interaction and exploration. Systems may have either been designed as

stand-alone at one extreme, or interoperability may have been planned and designed into their infrastructure. The extent to which this has taken place will affect the potential for building seamless discovery services.

From the point of view of an end-user, at least three increasingly complex levels of discovery service can be described, within which interactive exploratory behaviour can take place:

Level 1: Services underpinned by integrated discovery mechanisms.

This level relies on a well-known data format, well-structured metadata, key elements of chemical (and other) description, and specified search fields. The approach relies on tightly integrated services, designed to work together and exchange information. The parameters within which search can take place are known and clearly identified, and they are matched with the data that is being searched. This is the path that the eBank project has followed, by seeking to integrate within the digital library infrastructure through well-agreed protocols and structured metadata. At this level, only some aspects of the underlying data are represented and used in the discovery process, and the search may be limited to specific characteristics of data files, possibly in a well-known format. For example at this level, only data available in CIF format would be encompassed, as this provides a level of conformity that is likely to be adhered to by a wide variety of crystallography services, irrespective of their differing working practices. From the user's point of view, the search would be restricted to fields that are supported by the metadata. Due to the agreement between the participating parties, and the careful choosing of metadata fields, it would almost be guaranteed that the required specialised search fields would be made available, and that they would represent the content of the data accurately and consistently across the Federation partners.

Level 2: Access to all underlying data through data and text mining

This approach would widen the scope of the search carried out in Level 1 so that the metadata and search parameters available to the user would no longer be restricted to the metadata generated or submitted according to the agreed profiles. Additional information parameters on which to search could potentially be generated through automated methods, relying on text or data mining, and the connection between the data results and the queries would extend into other file formats. Thus the end user would see search results that are derived from a wider range of sources, both data files or text, and a degree of further processing (beyond the pre-determined selection of metadata) would have been applied to calculate the relevance of results. The system would not be limited by an agreed infrastructure and metadata profile, since it would be capable of accommodating formats and files with indirect links to the metadata, and would be able to search fields which are not necessarily represented by metadata. However an element of selection would be applied to the scope for the search, since only selected resources related to the partner infrastructure would be connected and linked in.

Level 3: Services that are all-encompassing.

This highest level of service would aim to offer to the user a very wide-ranging landscape of resources, which can be described broadly as "*everything that is out there*". In other words, any resource which has any relevance to a crystallographic search would fall within the scanned horizons. Such resources could be located or referred to within a very wide and inclusive vista, and would not require any prior agreements such as those defined by the scope within a Federation. With such an imagined landscape, it is clear that no integrated infrastructure would exist, or rather, the agreed infrastructure would only extend as far as any agreements reach. By its very nature the characteristics of the resources are vast and heterogenous to the extreme, and beyond the reaches of the Federation, any type of system infrastructure could be encountered. Although aiming to be all-encompassing, some obvious restrictions, for example controlled access, would pose limits to the scope. However the aim would be to include anything and everything that is marked as "relevant". For the end user, an element of serendipity would come into play since encounters with unexpected sources or surprising connections would be more likely to appear. However in this scenario the searches are much less controlled, and the results are less well-selected.

In this scoping study, the focus is on achieving a Level 1 type of service, i.e. integrated technical infrastructure, building on the cumulative experience of the eBank and SPECTRa projects. However Level 2 and 3 services are also reviewed, and placed in context.

The Federation will consist of a complex network of organisations taking on the roles of owners (producers) or custodians of information (data, literature or other resources), and service providers managing and facilitating access to those resources. Owners and custodians produce and store resources in order for the resources to become accessible for end users either immediately or in the future. Service providers manage interactions between end users and those resources, through discovery services, additional processing (e.g. visualisation) and where required, managing access rights. One organisation may take on multiple roles, or offer a variety of services. Although the fully-functioning Federation is some way ahead, some technical components for storage and management of resources can be identified within the existing infrastructure, and are reviewed in this section. An assumption is made that future Federation infrastructure will need to incorporate such existing solutions to some extent, rather than starting from scratch.

5.7.2 Solutions and experience from the digital library community.

5.7.2.1 The Z39.50 protocol

Z39.50 is a standard for information retrieval maintained by the Z39.50 Maintenance Agency administered by the Library of Congress, with a development history dating back to the 1970s. It is a protocol which specifies data structures and interchange rules that allow a client machine (called an 'origin' in the standard), to search databases on a server machine (called a target in the standard), and retrieve records that are identified as a result of such a search. The protocol defines interactions between two machines; applications can be built on top of the protocol to manage concurrent connections to multiple distributed machines. The protocol is perceived to have limitations[44] and has not been pursued further within the Federation.

5.7.2.2 The Open Archives Initiative for Metadata Harvesting Protocol (OAI-PMH).

OAI-PMH is a protocol which forms part of the Open Archives Initiative which "develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content." The detailed specification of Version 2 is available[45] and has been used to underpin the exposing of repository content within the wider information environment. Various harvesting services have been built based on this foundation, including the PerX pilot service in the engineering domain.

5.7.2.3 The PerX project

The PerX project[46] developed a pilot service which provided subject resource discovery across a series of repositories of interest to the engineering learning and research communities. This pilot was used as a test-bed to explore the practical issues that would be encountered when considering the possibility of full scale subject resource discovery services. There are lessons to be learnt from the PerX Project when considering an infrastructure that is dependant on the OAI-PMH protocols and data providers:

"metadata providers rarely follow the OAI-PMH standards and recommendations in full, and also that commercial content providers often have little interest in OAI-PMH."

"much of the metadata produced by data providers contains errors and omissions which can cause problems for service providers, or, at worst case scenarios, make the metadata unusable. Largely because of this, and despite some time being spent on various attempts, it proved impossible to automate the reharvesting process to any degree. There are many limitations with the OAI-PMH approach." "successful ongoing maintenance of OAI targets, for example, would require a mixture of automated and manual approaches and that the level of ongoing maintenance required for OAI targets in a live service would be relatively high."

5.7.2.4 OAI-ORE Object Re-Use and Exchange

The Open Archives Initiative has secured funding from the Mellon Foundation and Microsoft to develop further specifications that "allow distributed repositories to exchange information about

their constituent digital objects. These specifications will include approaches for representing digital objects and repository services that facilitate access and ingest of these representations. The specifications will enable a new generation of cross-repository services that leverage the intrinsic value of digital objects beyond the borders of hosting repositories". The specifications are in their third alpha release[47] and "describe a data model to identify and describe aggregations of web resources, and the encoding of the data model in the XML-based Atom syndication format". The eCrystals team at Southampton is working with the OAI-ORE team to use the crystallography domain as a test-bed for implementing OAI-ORE as part of the Microsoft-funded eChemistry Project.

5.7.3 Interactions with third party services

The context of the Federation is defined by the collaborating parties, and any potential agreements for interaction achieved amongst them. However it would be short-sighted to ignore that the Federation operates within a wider environment, and the information landscape consists of organisations and services outside of the Federation, with which the Federation may wish to interact. Additionally third parties may wish to access the resources within the Federation and present them within other, independently managed, contexts. This wider information landscape is now considered.

The eBank Project has identified a number of services that form part of the online information landscape, all of which are of potential relevance to Federation interactions. Some of these fit within digital library infrastructures, employing, for example, protocols such as the OAI-PMH. Others are independent services not designed to fit within specific frameworks. The range extends from generic, all-purpose services (such as OAISTER, which encompasses OAI-PMH-compliant sources across the spectrum of disciplines and resource types (images, sound, publications, web pages) to crystallography-specific interfaces (such as the COD, which contains user-contributed data, or ChemRefer, which specialises in searching chemistry literature). More detail is now provided about these identified services to illustrate the breadth that a Level 3 discovery service would imply.

5.7.3.1 Crystallographic Data Aggregators

eBank-UK Aggregator http://ebank.ukoln.ac.uk/

The eBank UK Project has implemented a demonstrator of an aggregator service which consisted of the metadata from the eCrystals repository at Southampton and a sample of metadata on publications provided by IUCr. The aggregator demonstrator has been used to explore aggregation issues for a discovery service based on metadata. It also shows how the metadata exported by the Southampton eCrystals archive could be used by a third party to provide search services. The exported metadata supports the following search fields:

Title - name by which the resource is formally known

Creator - Creator(s) of the dataholding

Subject - Keywords selected from an adapted version of the IUCr World Directory of Crystallographers list

Subject - Chemical compound, identified with an InChI (International Chemical Identifier)

Subject - Chemical compound, identified with a Chemical Formula

Subject - Chemical category: CompoundClass

Publisher - Affiliation of the creator(s)

Modified - Date on which the dataholding was changed

Created - Date of creation of the data holding

Type - All dataholdings are given the type: "Crystal structure data holding" as a fixed value

Identifier - An unambiguous reference to the resource within a given context: This is the Crystal Structure Report URL

Identifier - An unambiguous reference to the resource within a given context and has the following syntax: DOI:10.1594/ecrystals.chem.soton.ac.uk/[acronym/number of our choice]

Has Part - References to data files which are part of the data holding.

Rights - URL pointing to a general plain text rights statement

The Cambridge Crystallographic Data Centre CCDC (http://www.ccdc.cam.ac.uk/)

CCDC acts as an aggregator and subject repository, the CSD, for crystal structures containing organic moieties. To achieve this CCDC works in close collaboration with publishers and peer reviewers to generate its content, which ensures a collection that is validated to a high level of accuracy. Through necessity this work relies on a considerable amount of human intervention, which has the added curation benefits. CCDC has arrangements with virtually all journals containing crystallographic data whereby authors are required to interact with the database during the publication process.

The Chemical Database Service CDS (http://cds.dl.ac.uk/)

This service aggregates the crystallographic databases of the published literature, by offering a single interface to the licensed databases: CSD, ICSD and CrystMet. Value is added by additionally making technical details of the data searchable (eg space group or unit cell) and there is the intention to include more recognised sources of data as they become available.

CrystalEye (http://wwmm.ch.cam.ac.uk/crystaleye/)

CrystalEye is a crystallographic aggregator developed at the University of Cambridge, that crawls specific Websites for openly available data. These are currently predominantly publishers Websites, but the intention is to include repository data as it becomes available. The CrystalEye service indexes the molecules it has found and makes them searchable in a different manner that is useful to the practising crystallographer i.e. classification on geometric parameters (eg bond length) in addition to standard bibliographic terms. It would be preferable for repositories to provide such services with structured metadata that provides explicit context for the data, rather than relying on conventional Web crawling and inferring.

Chemrefer (http://www.chemrefer.com)

This service provides access to full text chemical, pharmaceutical literature Index through a simple, Google-like, search interface. It relies on locating already-existing sources, e.g. publications on individual's Websites, and improving search and access to them.

Crystallography Open Database COD (www.crystallography.net)

This initiative is similar in principle to the eBank and SPECTRa efforts in that it promotes open data. It allows individuals to submit data files to the service, either directly, or by reference (defining a 'REF' format for the latter). It contains over 40K entries, thereby representing a substantial source of data files. Where this effort differs is in its assumption that access to the data and searching will occur through the Web interface provided at the URL above. The service has not been designed to be part of a larger network or infrastructure that exchanges metadata.

5.7.3.2 Blogs, Wikis and Social Networking Sites

In the last few years, the use of blogs for publicising information and opinions on the internet has grown significantly and there are now a number that contain information or chronicles relating to the chemistry domain. However the majority of these are confined to the discussion of chemistry matters or the airing of opinions and relatively few are concerned with the technical details relating to particular experiments. UsefulChem[48], ChemTools[49] and OpenWetWare[50] are blogs and wiki-based services that have been built to enable Open Notebook Science[51] and have the ability to support data from scientific experiments. Whilst this support is currently somewhat rudimentary, there is interest in developing such sites for the purpose of discussion of scientific experimental data, including crystallography and the integration of tools is likely to follow. Blogs tend to support RSS feeds to push their content out to the Web but do not support

more complex Digital Library protocols, however they provide an ideal framework in which to develop concepts such as SWORD[52] and OAI-ORE

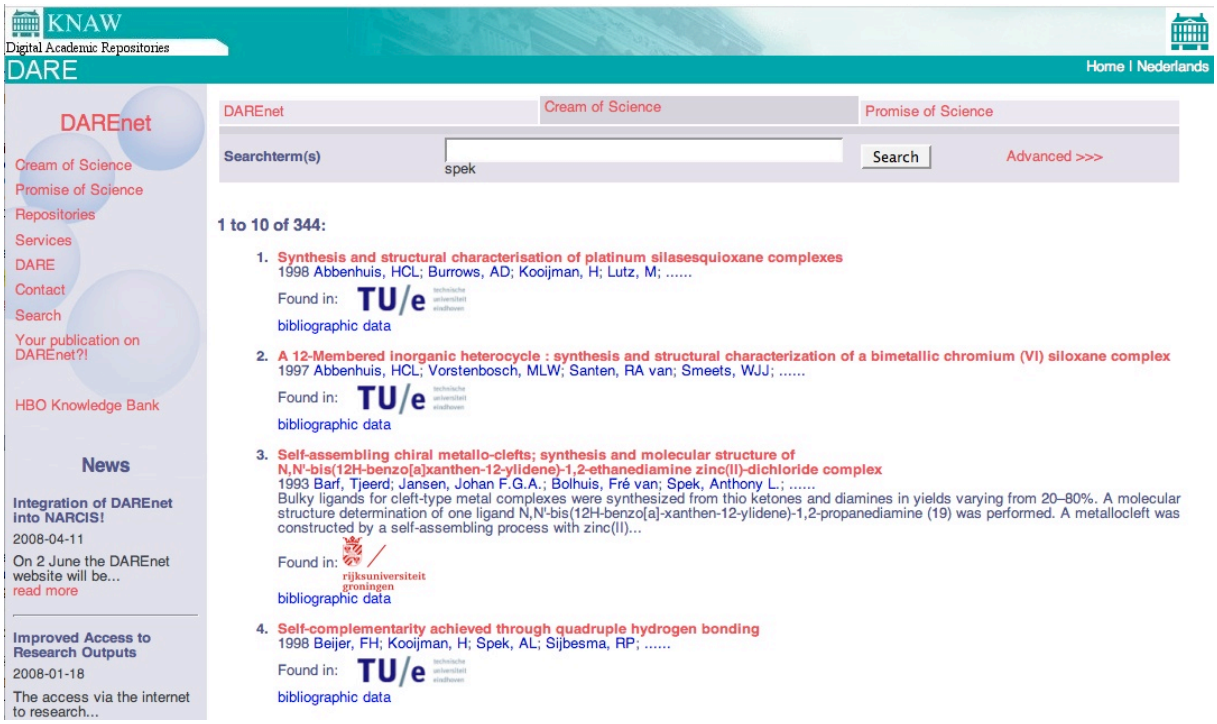
Other Web 2.0 resources currently hold less chemistry content, however the notion of Virtual Research Environments to support certain aspects of this field is being investigated. The myExperiment Project[53] has constructed a VRE based on the model of a social networking site where groups can exchange and discuss experimental data either openly or in a closed manner. Being purpose built, this resource will be relatively easily able to support tools and protocols for handling and disseminating experimental data.

5.7.3.3 OAI-PMH harvesters containing crystallography information

DAREnet (www.darenet.nl)

The DAREnet service is based in the Netherlands and aims to provide "Worldwide access to Dutch academic research results". Harvesting a network of OAI-PMH data providers based at Dutch academic institutions, it makes available a search interface to access the aggregated metadata. The interface is available as a simple search and as an advanced search; however the advanced search (predictably) is aimed at general searching and is not specific to crystallography (e.g. it does not search any specific chemical values or terms). Use of these interfaces show that the collection does contain results (mainly publications) which would be relevant to a crystallography search: using the term "crystallography" yielded 40 results, many similar to the ones included in the eBank demonstrator.

The general advanced search supports fields such as author and year, as shown below in Figure 5.



OAIster (http://www.oaister.org/)

Like DAREnet, this is an example of an OAI-PMH aggregator which is even more wide-ranging and inclusive: it encompasses any repository and all content types, harvesting metadata from 675 institutions. This aggregated metadata is searchable, and searches can be limited by resource type. Search results are pointers to collections of data and there are five types of dataset found across several resources that give 2000+ results for a 'crystallography' search term.

These aggregator services are summarised in Table 1 below.

	Aggregator Service	
	Generic	**Crystallographic**
Data and Literature	OAIster	
Data only		eBank SPECTRa ReciprocalNET CrystalEye
Textual publications only	DARENET	ChemRefer

5.7.4 Evaluation of architectural options.

Level 1 Services

From the combined experiences of eBank and SPECTRa projects, it can be seen that local practices will impact on the metadata that can be supplied, the tools used for collecting that metadata and that the approach requires tight agreement and conformance. It is not yet known if this is feasible within the crystallography domain, and to date there are few instances of OAI-PMH compliant crystallography-specific services. There is the advantage that Federation partners are willing to collaborate, however the PerX Project reports reluctance on commercial providers to provide OAI-PMH targets, and therefore realistic expectations have to be held with regard to the inclusiveness of any service that requires tight integration and conformance by data providers. At present, there is a lack of testing of aggregation issues due to lack of availability of metadata in critical mass proportions. This aspect will be revisited in Section 5.12.

Level 2 Services

These require Level 1 service infrastructure to be in place. Within that infrastructure consideration needs to be given as to how to refer to the data sources that can then be mined e.g. the granularity of identification within descriptions, restrictions to access (embargoed content). The requirements for Level 2 need to be kept in mind when designing Level 1 services, including terminology issues which may need to be addressed prior to data-mining (e.g. dictionaries). Co-operation with bodies such as NaCTeM is essential.

Level 3 Services

For the purposes of the scoping study, the third level of service described in Section 5.7.1 can be viewed as not being immediately achievable. eBank has set the groundwork for achieving Level 1 service, albeit work is still required to advance even that Level 1 option. However it is desirable that whilst aiming to achieve level 1, the more exciting possibilities of a Level 3 service are explored, for example through scenario-building. The Level 1 technical infrastructure can form the basis on which to think about the issues for a Level 3 type service. The outside services identified in this report are all relevant, however the list is non-exhaustive. Competition may exist from other generic type services e.g. Google, which can search InChIs, so the Federation needs to think of unique features which would give it a competitive advantage.

From an architectural point of view, eBank has chosen the route of tight integration of services through agreed protocols and metadata description. More experience is needed in this area to validate the approach as a basis for the integrated approach to discovery. Challenges may

exist in gaining enough critical mass and participation and it is clear that it would not provide the complete level of discovery service required: therefore Level 2 and Level 3 type services need to be considered as the longer-term aim. The techniques required for those two approaches have not been part of the eBank Project to date and more knowledge and experimentation is also required in that area.

5.8 Rights and Licensing

The study raised a number of rights issues associated with raw and derived data. The rights associated with the raw and derived data from a single experiment, may be different. Third party services such as the large scale ISIS facilities at STFC, may wish to keep the raw data, but allow users (visiting scientists) to "take" the derived dataset. Ownership of the IPR in this case should be clearly stated at the outset and indicated in the metadata in the repository. A number of publication scenarios may be envisaged, which manage the release of datasets in line with third party policy (e.g. NCS and STFC/ISIS). Repositories provide the framework to attribute IPR at the time when the data is generated and captured. This should reduce the amount of 'lost' data and avoid blocks on publication, such as the author moving institution. The University of Southampton is considering the legal repercussions of data loss and CCDC have also taken legal advice on this point.

One mechanism to manage the public dissemination of data is through embargoes. Repositories facilitate the management and implementation of embargoes, and this is an effective way of re-assuring researchers that their data is being made available more widely in a managed environment. It is to be hoped that attitudes towards open data will change and that standard embargo mechanisms within repositories will become the norm. The NCS policy statement for data publishing[54] states that there is a three-year embargo for unpublished structures, which is open for consultation, and provides a method for dealing with all unpublished data. When data is released for publication in the eCrystals repository, the structure moves from the private to the public archive and a citation is produced. A contrasting approach has been implemented by the SPECTRa Project, which has embargo "buttons" for selection by the depositor.

As a further example, the National Crystallography Service (NCS) is based within the Department of Chemistry at the University of Southampton. The NCS has formulated a policy for embargo implementation, and it is notable that this embargo has not been set arbitarily: three years has been set as the embargo period, which is the length of a PhD or average research grant. The embargo mechanism can be implemented automatically, i.e. implemented without a warning notice, but it is also possible to set a notification to be sent to the chemist one week before the data publication date, and the data record can be reviewed again at that point.

Some journals are very concerned about "prior publication" of data or other research outputs in a repository. It is not always clear exactly what is meant by "prior publication". What is the difference between prior publication in a repository and replicating (some of) that data and/or information in a peer-reviewed paper? What proportion of a paper should be primary data (which is also in a repository)? What is the position with regard to derived or processed data and raw data in this context? How can the application of licences support open data?

It is clear that well-defined rights information should be contained within the metadata schema referencing appropriate licences and data-sharing protocols. There are a number of licensing initiatives which are developing protocols for open data including the Science Commons CCZero[55] Framework and the Open Data Commons Public Domain Dedication and Licence (PDDL)[56] now available as a beta Draft release. More work is needed to explore their implementation within the Federation of crystal data repositories and more widely in other disciplines within the UK research arena.

Recommendation 4: JISC should seek expert advice to advocate the implementation of appropriate open data licences to provide a common basis for data sharing within the research community.

5.9 Data Quality and Validation

The quality of data deposited in institutional repositories via the self-deposit model, is a cause of concern to both the Cambridge Crystallographic Data Centre and chemistry publishers. There was also a view that funder mandates to deposit data within institutional repositories rather than in domain data centres, could cause a reduction in the quality of data/records. Repository policies need to be formulated to address this issue. In addition, the quality of a dataset held in a repository needs to be clearly indicated and demonstrated to the (re)user. How can the quality of a crystal structure be measured and assigned? Can this process be achieved automatically without human intervention? There was a view that the data and metadata validation and quality assurance process should be initiated at the point of data generation, and it should be noted that validation is required for both the dataset and the accompanying descriptive metadata. Furthermore, QA processes need to be applied across all Federation partners. For this to be achieved, validation mechanisms need to be established as part of the appraisal process, and policy statements for an institutional repository should contain clauses relating to the QA procedures.

There are other approaches to quality indicators. In some areas of bio-informatics, the data must be deposited in an appropriate database such as those at the European Bio-informatics Institute before publication of the article. In this way, deposit of protein sequences in one of the UniProt databases provides a data validation mechanism prior to publication. Despite such efforts, as scientific fields evolve the requirements and use of data changes and this effect has been observed by the Protein Databank who have had to undergo an extensive and costly remediation project[57] to correct data holdings and bring them in line with new approaches to using them. In general, datasets are not peer-reviewed before publication, and there are currently no formal or agreed processes in place to facilitate this approach. Facilitating access to datasets by referees as part of the peer-review process is being actively explored by the eCrystals Federation. The application of open standards for identifying individuals is very relevant and the OpenID[58] initiative is relevant in this context. An alternative method is the open and collaborative approach (rate my data!) adopted by sites such as Swivel[59] and Many Eyes[60], which use social networking methods to rate datasets.

The JISC, RIN and NERC have jointly commissioned a study to investigate the publication and quality assurance of research data outputs. This study, which has been carried out by Key Perspectives, is due to report in 2008, and it is hoped will provide some guidance for the Federation and other repository networks. Requirements for data quality may also be influenced by elements of the new UK Research Excellence Framework (REF)[61]. In the context of crystallography data, crystal structure results published in *Acta Cryst. E* provides a citation and impact factor but currently institutional repository records do not. The REF has the potential to provide incentives and drivers for data self-deposit in institutional repositories and to some extent, to change the research culture in the UK with regard to data sharing.

Recommendation 5: JISC should consider funding further work to support data validation and data quality assurance methodologies, possibly taking a domain-centric approach.

5.10 Preservation, Curation and Sustainability

Responsibility for the long-term preservation of crystal structure data is unclear. ReciprocalNet does not have a preservation policy or model. The possibility of applying a LOCKSS[62]-type model was raised during the interviews. Responsibility for provision of a preservation service was also raised by the Royal Society of Chemistry, and they would like to see assurance that data that they would no longer hold, would be stored in repositories, and would be managed and available in the longer term.

CCDC and IUCr are currently the major organisations in the field. The CCDC subject repository has an established history (forty years old), but surprisingly, has no formal preservation policy. There is a “gentleman’s agreement” that IUCr is the alternate store for CCDC data, but there is no formal agreement in place. The view taken is that since CCDC have been operational for forty years, the arrangement works. Clearly however, there are some substantial risks

associated with this premise. It was also observed that the community acts as a backup for the datasets to some degree, since the crystallography laboratory community is active and searches the database on a daily basis. There is a news group who provide user feedback to correct records. This represents an early example of a wiki-type model. Direct contacts from other crystallographers provide a community editorial role – an exemplar of the community curation model, where there is active checking for accuracy and integrity, without reviewing scientific interpretations. In one sense, community use of data guarantees its longevity and the absence of use of data, raises questions of retention: put simply, if a dataset is used, then it is important for someone. However, this informal approach does not provide the sustainability required for assured long-term preservation of the data.

There is a belief that not all data should be stored for the longer term and some clarification of the appraisal process is required. Much of the determination of structural models to fit the data (best fit) is software-driven. Whilst it is evident that software will improve with time, how do you know what data to keep, in case future more sophisticated software can model it more effectively? Some assessment of the viability of quantitative classification of the data in the archive is required to enable machine parsing of quantitative criteria for selection or de-selection of datasets. The development of such quantitative criteria for appraisal, validation and goodness-of-fit for structural modelling matches, is recommended. Appraisal criteria might include consideration of whether an experiment is repeatable and at what cost. Many experiments are opportunistic and chemists carry out as much analysis and processing of the sample and data as they can, at the time of synthesis. The development of criteria to define reproducible experiments, would be valuable.

There are also preservation and sustainability constraints associated with the proprietary binary formats generated from instrumentation. In general, the instrument manufacturers do not provide any indication of the sustainability of their software, but this forms a critical element of the representation information to be collected based on the OAIS Reference model, which underpins many operational repositories and archives. There is a need to "future proof" datasets for potential analysis by new algorithms that may be developed by instrument manufacturers in the future. The raw processing of data is largely done by very proprietary bespoke software, and compilers written by the scientist, so there are additional issues around preservation of the software associated with this practice. All of this information should be captured in the representation information.

In contrast, STFC keep a record of every experiment carried out and that data has been successfully migrated across platforms and formats. However there is no curatorial function, merely a migration of formats. There are perceived cost-benefits in curating the data to minimise future repeated experimentation. The STFC ISIS facility has no formal preservation policy but this is under active consideration in the R&D section. The departmental nature of STFC raises questions about strategy and policy across the whole organisation (and many other institutions which operate on a departmental basis).

Data from Diamond experiments on beam lines goes to an intermediate store and then is transferred into the Atlas Datastore for long term storage. There are issues around who manages this process? Instrument scientists may not know that the scientist working remotely has completed their experiments, and there are workflow issues and complexities of managing data in such highly distributed and fragmented procedures. The workflow needs to be well-defined and include consideration of curation and preservation elements, to ensure effective data management. The economic sustainability of large-scale facilities like Diamond, is intricately linked to future predictions on data growth and scale-up, in relation to the costs of data storage, as evidenced by the STFC Delivery Plan 2008/9-2011/12[63].

The preservation requirements of large-scale science are rather different to those requirements of the highly fragmented and diverse small science activities carried out in a multitude of labs around the world: the so-called "long tail science described by Jim Downing and blogged by Peter Murray-Rust"[64]. It has been observed that "*small science is horribly heterogeneous and far more vast. In time small science will generate 2-3 times more data than big science*".[65] However, the sustainability cost model for data preservation services such as the Atlas Petabyte Datastore developed to support outputs from large scale facilities, may not be

appropriate to meet institutional laboratory data requirements. In addition, there is considerable sub-disciplinary variation within the chemistry domain in terms of the scientists' views on the importance of keeping data, ranging from capturing everything at source, to a rather more random and unmanaged approach to data preservation. A better approach is to develop a Data Management Plan for a particular experimental technique i.e. address the issue at the process level of granularity, since a single facility would encompass many experimental techniques.

In addition to data from which published articles are derived, there are also preservation issues associated with chemistry data accompanying theses. These have been investigated by the SPECTRa-T Project[66] team and prospective partners in the Federation. There is a fundamental question to be asked if data is present as an integral part of a thesis: do you store data with the thesis or not? DSpace has not been designed to connect repository instances, so a third party service may be required to achieve this, or some other technical solution. In such cases, the depositor needs to be aware of the limitations of local or different repository platforms and architectures. This should be a part of the advocacy given by the Library or repository manager and should be included in any preservation policy. There need to be clear policy statements describing what a repository does do and what it doesn't do in terms of functionality and preservation capability. The scientist or user needs to know the economic sustainability guarantees underpinning the repository, i.e. project-based, departmental, institutional, funder, national library etc. However, this vital information needs to be communicated to the depositor quickly and effectively, and ideally not buried in multi-page policy documents. Such policy needs to be considered from a Federation point of view in terms of parity and commonality across institutional partners: it is highly likely that partner institutions will vary in their practice in managing theses and accompanying datasets. There is scope for the JISC/CURL-funded EThOSnet Project[67] to begin to address these issues at a national level.

Further eBank work on preservation requirements of data repositories is presented in a separate Report "*A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed federation*".[68]

Recommendation 6: JISC should fund the development of quantitative criteria for the appraisal of datasets. These criteria should take into account how the reproducibility of an experiment can be described in a "standard" manner.

5.11 Community and Inter-disciplinary Interactions

The final two sections focus on broader issues and the first describes lessons learnt which will enable leveraging of interactions within and beyond crystallography (and beyond chemistry).

The strengths and weaknesses of the domain approach which have emerged through this scoping study highlight lessons for other disciplinary communities seeking to develop institutional Federations or other networks of repositories or data archives. We emphasise the critical importance of the following "Checklist of Community Criteria for Interoperability" Table 2.

Community Criteria for Interoperability	Crystallography exemplars
1. Involvement of professional bodies and publishers.	Royal Society of Chemistry, IUCr.
2. Development and adoption of a common domain data format standard.	CIF
3. An established data validation mechanism.	CheckCIF.
4. Implementation and adoption of a common domain identifier.	InChI
5. A metadata schema application profile which supplies	eBank-UK schema

a common core element set.	
6. An existing subject repository, which may operate on a commercial basis.	CCDC
7. A degree of homogeneity and co-ordination in disciplinary research practice.	CIF and COMCIFS
8. An established service ethic and associated policies, which drives research practice for the common good.	NCS or CCDC or CDS

Table 2 Checklist of Community Criteria for Interoperability

However, there are also a number of parallel constraints, barriers and "Disruptive Effects" which work against the Community Criteria, create tensions and conflict, and ultimately inhibit creative inter-disciplinary interactions. Clearly, to stifle or remove diversity and innovation in the research context would defeat the purpose, however there is a subtle balance between advocating and facilitating common practice and "to let a thousand flowers bloom".

Some Disruptive Effects are listed in Table 3.

Disruptive Effects	**Mitigating Action**
1. Diversity of internal laboratory practice and culture.	Best practice standards, advocacy, core standard formats, AP
2. Arbitrary re-use of data because of "lock-in" to instrumentation and proprietary software e.g. CSD.	Advocacy, core standard formats, AP
3. Data re-use is limited because only processed (not raw) data is shared more widely.	Capture and expose raw data in laboratory repositories.
4. Limited data-sharing culture within crystallography, which inhibits wider chem-informatics.	Advocacy, awareness-raising, tool development
5. Inter-disciplinary re-use of data depends largely on human interaction and is hindered by lack of m2m interfaces.	Develop Web services such as CrystalEye which operate across distributed repositories.
6. Formal publishing disconnects inhibit interdisciplinary interactions e.g. lack of embedded links between domain identifiers such as LSIDs and InChIs.	Advocacy, awareness-raising, and partnerships with publishers. Develop knowledge extraction tools
7. Competitive relationships between institutions, departments and laboratories, as a result of research assessment frameworks and funding awards.	Consortium agreements should include clauses on data-sharing.
8. High-level strategic fragmentation associated with data management plans within and between the funding bodies.	Co-ordinated strategic planning for data curation *across* research councils, other funders.

5.12 Collective Intelligence and Open Science.

We are seeing a growing momentum behind the concepts and practice of Open Science, enabled by collaborative technologies such as blogs, wikis and social networks, where

scientists can comment, voice opinions, develop ideas, produce grant proposals[69], share methodologies (OpenWetWare) and post results[70]. Repositories are positioned within this fluid space and have the potential to provide robust infrastructural foundations for a critical mass of open, reusable scholarly content. This content may include raw, processed and derived data.

A complex Web of cross-links, cross-references, co-citations, cross-posts, discussion and back channel chat is emerging, some of which is focussed on repository content supported by annotations, tags, ratings and votes. Of course this "evidence" provides a wealth of additional information for consumers, but there is a growing challenge in finding, viewing, organising, sorting, filtering and generally managing this "deluge of discourse". Aggregator services such as Google Reader can be used for this purpose, however in general we have not moved beyond considering these data clumps and data clusters as simple aggregations. We need to move forwards in our thinking and view this primary material as "collective intelligence" which needs to be actively curated, packaged in digests and rendered in visualisations. The data elements can be assessed, manipulated in models and other secondary forms, analysed for innovative trends and mined for new knowledge. The development of interactions between repository content, repository services and this collective intelligence are as yet relatively rudimentary, but the potential for this resource to enhance and enable new and exciting open science is significant.

Recommendation 7: JISC should fund a scoping study to investigate the potential of collaborative technologies, collective intelligence and repository content and services, to stimulate new modes of open science.

6 Appendix

6.1 Interview Pro-forma

1) Coordination and advocacy

a) Technology employed - how has this decision been influenced?

b) Does the deposition process require assistance from / mediated by an expert?

c) What levels of advocacy have been required in order to get people to deposit?

d) What incentives have been provided / mandates employed to get users to deposit?

e) Do researchers see the conventional publication process as conflicting with depositing in a repository?

f) What kind of services built on a federation of data repositories would provide depositors with an incentive?

g) What search / discovery / browse services would a researcher require at an individual repository level?

h) Does your repository comply with a known metadata standard or has the application profile been developed for this repository only?

i) Are you prepared to adhere & contribute to the eBank metadata application profile? Would this require guidance?

j) Do you see any barriers to choosing the eBank software?

k) Do you see a role for a subject repository in the federation model?

l) What are your views on using a subject repository, as opposed to a distributed federation of institutional data repositories?

2) Technical Interoperability & Standards

a) Is access management required / in place?

b) Is there a documented workflow for the deposition / ingest process?

c) When is deposit / ingest performed (integrated into workflow or as one process when experiment is complete)?

d) Is the workflow a standard process, same/similar to others or independent of deposition / ingest?

e) Have you documented your internal file schema (i.e. file types/formats)?

f) What are the number and complexity of file formats that can be supported?

g) Does the ingest process or dataholding / record make a distinction between raw and derived data?

h) Does your [Should a] repository contain raw (proprietary/binary) data or links to raw data?

i) How is metadata generated / captured (captured during workflow, extracted from deposited files or depositor keystrokes)?

j) Do you have any quality control over the completeness / validity of metadata generated?

k) Do you have any quality control / validation criteria for the dataset?

m) Are all records [immediately] public?

n) Is there an embargo mechanism / control over release into the public domain?

o) How are arrangements made with all concerned parties for a record to go public?

p) Should there be a standardisation across a federation on the format / layout and presentation of a record?

q) Can you provide representation information for your repository?

3) Semantic Interoperability & Standards

a) Should there be an eCrystals Federation application profile that all member repositories should adhere to?

b) How might a Federation application profile be enforced?

c) Are you agreeable to the eBank application profile being used as a working model for the federation?

d) Are you able to conform to the eCrystals application profile / trial it / contribute to it / provide feedback?

e) Alternatively should there be some form of centralised normalisation for metadata?

f) Do you use InChI in a record / in the OAI?

g) How do you generate / validate an InChI?

h) What generic persistent identifiers do you use / should be used?

i) Are you willing / able to pay for a persistent identifier / ability to resolve a persistent identifier?

j) What kind of role does cataloguing / terminology play in repository use / subject indexing / access features?

k) Do you employ rules / standardisation in titles (IUPAC nomenclature) / chemical formulae?

l) Does your repository contain data from different domains / techniques?

m) Do you see a use for using keywords to describe data files / data within files / context for data files? [process/context/object]

n) Is there a need to search within files or is context sufficient?

o) What should keywords express?

p) Would text mining be a suitable alternative to indexing by keywords?

q) How will researchers use repository data...will terminologies be required for:

i) Methods - entities / mining / knowledge generation

ii) Services - search / browse / harvest / data extraction

iii) Disciplines / users - bio, geo, eng, phys, chem

iv) Levels / roles - researcher / public / student

v) Record / document types - Data / data holding / publication / metadata

r) Is there a suitable terminology in existence or is it necessary to devise your own?

s) How might a terminology be maintained?

t) Is there a need for keywords to describe the experimental process?

u) Is there a need for keywords to describe context?

v) Is there a need for keywords to describe a digital object?

w) Is there a need to develop the keyword approach into an ontology?

4) Preservation & Curation

a) Do you have a mission statement regarding long term commitment?

b) Do you have a succession plan for when current funding ceases?

c) Have data files / types been documented / described for curation purposes?

d) Do you have plans for financial sustainability?

e) Is an aggregator service / subject repository harvesting from Institutional Repositories a feasible approach to preservation?

6.2 List of individuals participating in the interviews.

Organisation	**Names**
CCDC	Jenny Field, Robin Taylor, Frank Allen, Owen Johnson, Ian Bruno
CDS	Bob McMeeking, Don Parkin, Dave Fletcher
Chemistry Central	Bryan Vickery, Matthew Cockerill
IUCr	Brian McMahon, Peter Strickland
ReciprocalNet	John Huffman, Maren Pink, Kia Huffman
RSC	Richard Kidd, Colin Batchelor, Graham McCann
STFC	Brian Matthews, Shoab Sufi, Ken Shankland, Damian Flannery, Alun Ashton

University of Cambridge / Imperial College / SPECTRa	Peter Morgan, Peter Murray-Rust, Henry Rzepa, Alan Tonge
University of Sydney	Peter Turner
EBank-UK Project	Liz Lyon, Simon Coles, Mike Hursthouse, Jeremy Frey

Table 4 List of interview participants and organisations.

7 References

Links checked 2nd May 2008.

[1] Lyon E. 2003, eBank-UK: Building the links between research data, scholarly communications and learning. http://www.ariadne.ac.uk/issue36/lyon/

[2] Report from the eBank-UK joint workshop, London, October 2006. http://www.ukoln.ac.uk/projects/ebank-uk/workshop/eBank-SPECTRa-R4L-workshop/eBank-SPECTRa-R4L-workshop.pdf

[3] Project Prospect http://www.rsc.org/Publishing/Journals/ProjectProspect/

[4] Gene Ontology http://www.geneontology.org/

[5] Atmospheric Chemistry & Physics Public Review http://www.atmospheric-chemistry-and-physics.net/review/index.html

[6] OSCAR http://www.rsc.org/Publishing/ReSourCe/AuthorGuidelines/AuthoringTools/ExperimentalDataChecker/index.asp

[7] NeXus format http://www.nexusformat.org/Main_Page

[8] Claddier Project http://claddier.badc.ac.uk/trac

[9] JISC EnTag Project http://www.ukoln.ac.uk/projects/enhanced-tagging/

[10] SPECTRa-T Project http://www.lib.cam.ac.uk/spectra-t/

[11] Science Commons http://sciencecommons.org/

[12] Lyon, Liz. 2007 Dealing with Data Report http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/publications.html#2007-06-19

[13] RIN Data Stewardship Principles http://www.rin.ac.uk/new-data-stewardship

[14] NIH Policy http://publicaccess.nih.gov/

[15] Wellcome Policy http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/index.htm

[16] iRODS http://irods.sdsc.edu/index.php/IRODS:Data_Grids%2C_Digital_Libraries%2C_Persistent_Archives%2C_and_Real-time_Data_Systems

[17] ASPiS Project http://www.jisc.ac.uk/whatwedo/programmes/programme_einfrastructure/aspis.aspx

[18] LIMS http://en.wikipedia.org/wiki/Laboratory_Information_Management_System

[19] Barga, R.S, Andrews S. and Parastatidis (2007) A Virtual Research Environment (VRE) for BioScience Researchers. In Int. Conf on Advanced Engineering Computing and Applications in Sciences, pp31-38. http://www.ieeexplore.ieee.org/xpl/freeabs_all.jsp?isnumber=4401884&arnumber=4401895&count=16&index=10

[20] R4L Repository for the Laboratory Project http://r4l.eprints.org/

[21] MERCURY http://www.ccdc.cam.ac.uk/free_services/

[22] PublCIF http://journals.iucr.org/services/cif/publcif/

[23] ImgCIF http://www.iucr.org/iucr-top/cif/imgcif/index.html

[24] Diamond http://www.diamond.ac.uk

[25] http://www.synchrotron.vic.gov.au

[26] CIMA http://www.instrumentmiddleware.org

[27] *Acta Crys. F* http://journals.iucr.org/f/journalhomepage.html

[28] TARDIS Project http://www.tardis.edu.au/;

[29] CrystalGrid http://www.crystalgrid.org/

[30] SHELX http://shelx.uni-ac.gwdg.de/SHELX/

[31] CRYSTALS http://www.xtl.ox.ac.uk/crystals.html

[32] Olex Project http://sourceforge.net/projects/olex2

[33] eBank Metadata Application Profile http://www.ukoln.ac.uk/projects/ebank-uk/schemas/

[34] OAI-ORE http://www.openarchives.org/ore/

[35] Cheung, K. & Hunter J. 3rd Int DCC Conference 2007 presentation download from http://www.dcc.ac.uk/events/dcc-2007/programme/

[36] Microsoft Research eChemistry Project http://www.rsc.org/chemistryworld/News/2008/January/29010803.asp

[37] IEMSR http://www.ukoln.ac.uk/projects/iemsr/

[38] CrystalEye http://wwmm.ch.cam.ac.uk/crystaleye/

[39] Koch T. Report on Terminology and subject access issues, (2006) http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/termino-public.html

[40] Gene Ontology http://www.geneontology.org/

[41] EnTag Project http://www.ukoln.ac.uk/projects/enhanced-tagging/

[42] NaCTeM http://www.nactem.ac.uk/

[43] International Chemical Identifier (InChI) http://www.iupac.org/inchi/

[44] Lynch, C. The Z39.50 Information Retrieval Standard http://www.dlib.org/dlib/april97/04lynch.html

[45] OAI-PMH Version 2 http://www.openarchives.org/OAI/openarchivesprotocol.html

[46] PerX Project http://www.icbl.hw.ac.uk/perx/

[47] OAI-ORE Specifications http://www.openarchives.org/ore/0.3/toc

[48] UsefulChem http://usefulchem.blogspot.com/

[49] ChemTools http://chemtools.chem.soton.ac.uk/projects/blog/

[50] OpenWetWare http://openwetware.org/wiki/Main_Page

[51] Open Notebook Science http://drexel-coas-elearning.blogspot.com/2006/09/open-notebook-science.html

[52] SWORD Project http://www.ukoln.ac.uk/repositories/digirep/index/SWORD

[53] myExperiment http://www.myexperiment.org/

[54] NCS Policy for Data Publishing http://www.ncs.chem.soton.ac.uk/pub_pol.htm

[55] Creative Commons CCZero Framework http://wiki.creativecommons.org/Cczero

[56] Open Data Commons Public Domain Dedication & Licence
http://www.osbr.ca/ojs/index.php/osbr/article/view/516/475

[57] Remediation Project http://remediation.wwpdb.org/

[58] OpenID http://openid.net/

[59] Swivel http://www.swivel.com/

[60] Many Eyes http://services.alphaworks.ibm.com/manyeyes/home

[61] Research Excellence Framework http://www.hefce.ac.uk/research/assessment/reform/

[62] LOCKSS http://www.lockss.org/lockss/Home

[63] STFC Delivery Plan 2008/9-2011/12 http://www.scitech.ac.uk/About/Strat/Council/STFC_DelPLan.aspx

[64] http://wwmm.ch.cam.ac.uk/blogs/murrayrust/?p=938

[65] Carlson, S. Lost in a Sea of Science Data. The Chronicle of Higher Education 23/06/2006.
http://chronicle.com/free/v52/i42/42a03501.htm

[66] SPECTRA-T Project http://www.lib.cam.ac.uk/spectra-t/

[67] EthOSnet Project http://www.ethos.ac.uk/

[68] Manjula Patel, S. Coles, 2007. A Study of Curation and Preservation issues in the eCrystals Data Repository and proposed federation. http://www.ukoln.ac.uk/projects/ebank-uk/curation/eBank3-WP4-Report%20(Revised).pdf

[69] Science in the Open blog (Cameron Neylon) http://blog.openwetware.org/scienceintheopen/2007/12/12/the-open-research-network-proposal-update-and-reflections/

[70] Michael Barton blog http://www.michaelbarton.me.uk/2007/06/using-codon-adaptation/