Whilst institutional repository networks which provide managed storage and open access to the textual interpretations of research, are emerging e.g. SHERPA (http://www.sherpa.ac.uk/) and DRIVER (http://www.driver-repository.eu/), the data repository landscape within institutions is considerably less mature. Well-established community archives such as the UK Data Archive and the EBI sequence databanks for bio-informatics data, provide curated resources in certain disciplines. However, the technical infrastructure and associated support for research data remains fragmented and there are gaps in provision as exemplified by Open DOAR where out of the 76 recorded UK institutional repositories a mere 4 contain datasets. This is against a backdrop of an increasing "deluge" of data generated by both large-scale facilities and institution-based small-science. In addition, the highly social, participative, (and chaotic) constructs of the current Web environment are changing scholarly communications, and we are starting to see scientific data as well as textual information, being shared, discussed and evaluated in blogs and wikis e.g. within the associated R4L Project http://www.jisc.ac.uk/conference2007/. This is in contrast to the more formal standards-driven service-oriented architectural approach of the eFramework.

The pioneering JISC funded eBank-UK project (three phases since Sept 2003), has constructed an institutional repository that makes available the raw, derived and results data from a crystallographic experiment (http://ecrystals.chem.soton.ac.uk), developed the the eBank aggregator service for metadata harvesting by 3rd parties and promoted the linking from primary data to other research outputs within the scholarly knowledge cycle (Lyon, Ariadne July 2003). Phase 3 also investigated preservation and curation aspects of the data repository and evaluated approaches to audit and certification. Phase 3 was positioned as a transitional scoping study for the proposed eCrystals Federation, and this bid describes the first stages of full implementation. The results from Phase 3 are currently being assimilated and collated into a series of reports, however there are a number of outcomes which are already evident:

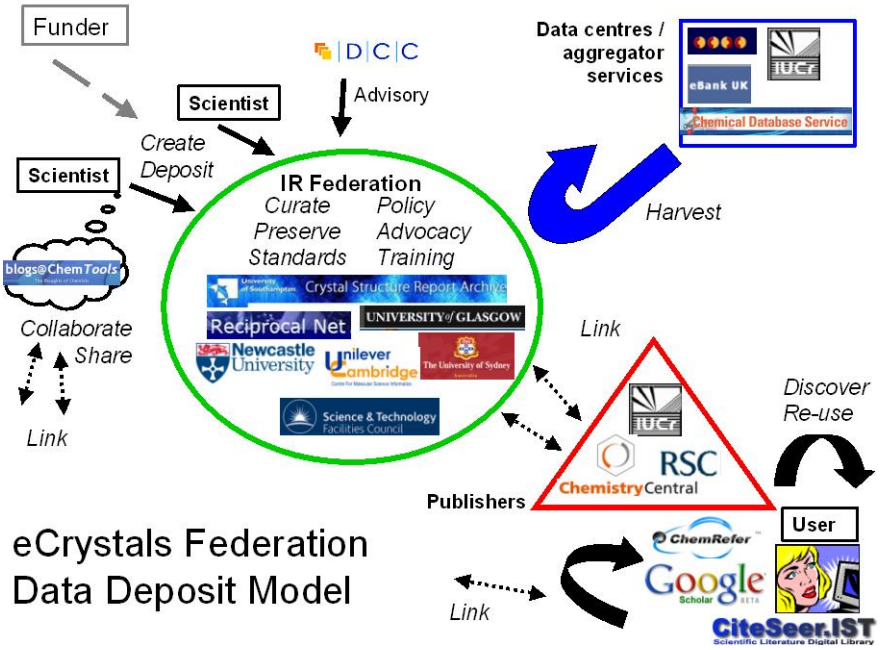
- Crystallographic laboratory practices are very varied, ranging from a more automated workflow with outputs handled and manipulated digitally, to a very "hands-on" process where an individual crystallographer oversees the process and maintains paper copies of results in a filing cabinet.
- This variation in laboratory practice has implications for the ease of adoption of a standard metadata schema such as the eBank Application Profile.
- Crystal structure data and associated information is complex, should be considered as compound objects and will require the use of a metadata packaging format such as METS or MPEG DIDL.
- There are likely to be a range of persistent identifiers in use within any discipline. The allocation of identifiers by the issuing agency must be efficient, reliable and scaleable.
- When considering preservation and curation, these aspects need to be addressed: audit and certification processes and procedures, representation information for crystallography data, preservation metadata for crystallography data, conformance to the OAIS Reference Model of repository software in use within the Federation.
- It is clear that preservation and curation issues will have to be addressed politically by both institutions and the community.
- Advocacy programmes will be essential to assist with populating the data repositories, since there is no established culture of sharing data within the chemistry domain.
- The implementation of a data embargo procedure/policy will be an important factor in encouraging searchers to deposit data destined for eventual open access.
- The pro-active support of professional societies, publishers, data centres and other key domain stakeholders is essential to achieve buy-in from the scholarly community.
- It is unclear as to the exact nature of the relationship between subject-based and institutional repositories and mechanisms for machine to machine interoperability will be necessary.

Building on these outcomes, the three key objectives of the eCrystals project are:

1. To create an operational Federation of data repositories in the crystallography domain thereby testing the effectiveness of the Federation Data Deposit Model, both within the crystallography community, and as a potential framework for other disciplines.

2. To make recommendations on preservation good practice for institutional data repositories
3. To assess the socio-political and technical interactions between subject and institutional repositories and sustainable models for partnership.

Figure 1: The eCrystals Federation Data Deposit Model.



In one sense, we can view the Federation as a form of Virtual Organisation (VO), which will develop its own set of governance agreements, policies, practice, behaviours, security and cultural values. It will be essential to develop these community aspects in parallel with the technical infrastructure.

Using the "roles" outlined in the *Dealing with Data Report* as a basis, the Federation partners can broadly be broken down into the following groups:

1) Institutions: Repository providers are the federation partners and have expressed their support (see Appendix). They comprise the Universities of Southampton, Cambridge, Glasgow, Newcastle, Indiana (USA, ReciprocalNet), Sydney and ARCHER (Australia) and STFC and represent institution-based repositories. The partners have been selected on the basis of their significance in crystallography, but also because they represent a truly global multi-platform data network.
2) Scientists: the individual crystallographers in the laboratory and practising chemists who create the crystal structures as part of their routine workflow.
3) Data centres: CCDC is a professional body with a subject repository for crystal data and CDS is a national service that provides federated searching across chemistry databases. They may be considered as the primary data harvesters of eCrystals.
4) Publishers: IUCr is the learned society representing crystallography, is a publisher of 8 journal titles and maintains standards for communicating and representing crystal structures. The RSC is a key publisher in the field and Chemistry Central is an emerging Open Access publisher who will operate a repository to store and link data relating to publications in their journals.
5) Users: scientists in related disciplines, students and other third parties who have a requirement to use crystallographic data as part of their research.

There are also two additional groups who are associated with the Federation:

6) Advisory services: the DCC will provide guidance on preservation and curation practice including the creation of preservation metadata and audit and certification tools. Institutional library and information services will play an important role in the

sustainability and preservation of repositories and will be engaged in policy matters from the outset (Cambridge and Southampton University Libraries leading). In addition IUCr has an interest in the preservation of federation data and scoping the operation of a subject repository.

7) Third party services: it is expected that third party services will develop across repository federation infrastructure and the model is being developed with this in mind. Whilst such service development will be the subject of future proposals, integration with the StORe project middleware and the CLADDIER Ping mechanism, which provide services to link data and publications, will be implemented as part of this proposal and the eCrystals Federation will act as a testbed.