eBank UK Linking Research Data, Scholarly Communication and Learning

Dr Liz Lyon[1], Dr Simon Coles[2], Dr Les Carr[3], Rachel Heery[1], Prof Mike Hursthouse[2], Christopher Gutteridge[3], Monica Duke[1], Dr. Jeremy Frey[2], and Prof Dave De Roure[3]

[1] UKOLN, University of Bath, UK
[2] School of Chemistry, University of Southampton, UK
[3] School of Electonics and Computer Science, University of Southampton, UK

Abstract. This paper presents an overview of the changing landscape of scholarly communication and describes outcomes from the innovative eBank UK project, which seeks to build links from e-research through to e-learning. As introduction, the scholarly knowledge cycle is described and the role of digital repositories and aggregator services in linking datasets from Grid-enabled projects to e-prints through to peer-reviewed articles as resources in portals and Learning Management Systems, are assessed. The development outcomes from the eBank UK project are presented including the distributed information architecture, requirements for common ontologies, data models, metadata schema, open linking technologies, provenance and workflows. Some emerging challenges for the future are presented in conclusion.

1 Introduction and context the scholarly knowledge cycle

The eBank project is predicated on the concept that research and learning processes are cyclical in nature, and that subsequent outputs which contribute to knowledge, are based on the continuous use and reuse of data and information [1]. We can start by examining the creation of original data, (which may be, for example, numerical data generated by an experiment or a survey, or alternatively images captured as part of a clinical study). This initial process is usually followed by one or more additional processes which might include aggregation of experimental data, selection of a particular data subset, repetition of a laboratory experiment, statistical analysis or modelling of a set of data, manipulation of a molecular structure, annotation of a diagram or editing of a digital image, and which in turn generate modified datasets. This newly-derived data is related to the original data and can be re-purposed through publication in a database, in a pre-print or in a peer-reviewed article. These secondary items may themselves be reused through a citation in a related paper, by a reference in a reading list or as an element within modular materials which form part of an undergraduate or postgraduate course. Clearly it will not always be appropriate to re-purpose the original data from an experiment or study, but it is evident that much research activity is derivative in nature.

The impact of Grid technologies and the huge amounts of data generated by Grid-enabled applications, suggest that in the future, (e-)science will be increasingly data-intensive and collaborative. This is exemplified in the bio-sciences where the growing outputs from genome sequencing work are stored in databases such as GenBank but require advanced computing tools for data mining and analysis. The UK Biotechnology and Biological Sciences Research Council (BBSRC) recently published a Ten Year Vision which describes this trend as "Towards predictive biology" and proposes that in the 21st Century, biology is becoming a more data-rich and quantitative science. The trend has clear implications for data/information management and curation procedures, and we can examine these further by returning to the concept of a scholarly knowledge cycle, figure 1.

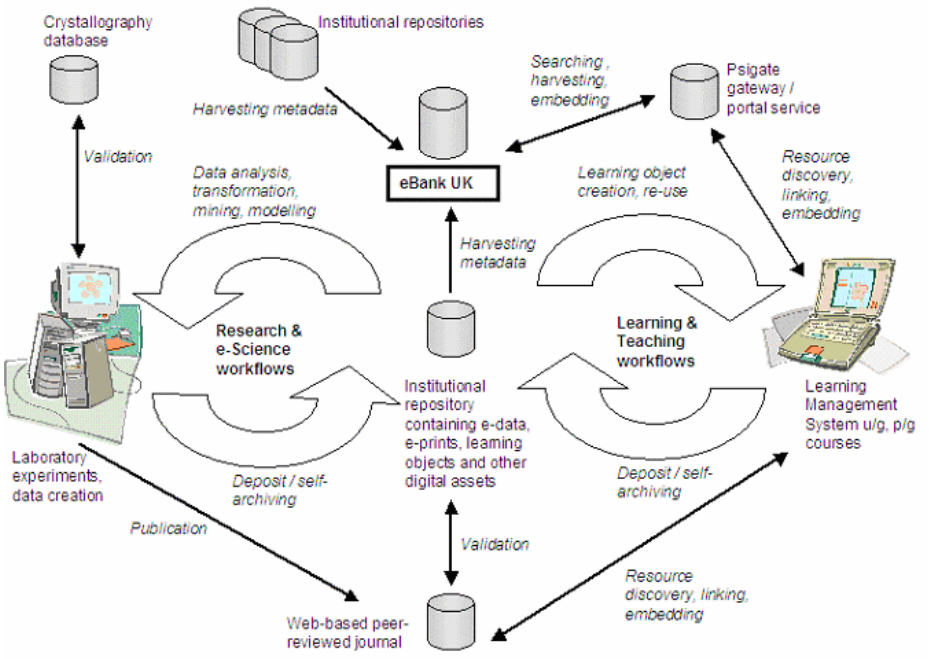

Fig. 1. Illustration of the scholarly knowledge cycle for research and teaching

A complete cycle may be implemented in either direction so for example, discrete research data could (ultimately) be explicitly referenced in some form of electronic learning and teaching materials. Alternatively, a student might wish to "rollback" to the original research data from a secondary information resource such as a published article or from an element within an online course delivered via a Learning Management System. In order to achieve this, a number

of assumptions must be made which relate largely to the discovery process but are also closely linked to the requirement for essential data curation procedures. The assumptions are:

- The integrity of the original data is maintained
- There is a shared understanding of the concept of provenance
- The original dataset is adequately described using a metadata description framework based on agreed standards
- A common ontology for the domain is understood
- Each dataset and derived data and information are uniquely identified (fig. 2)
- Open linking technology is applied to the original dataset and the derived data and information

In addition to the Grid computing context, these requirements relate directly to the vision of the Semantic Web, and Semantic Web technologies can be used to express the necessary relationships between objects. The application of Semantic Web technologies within the e-Science and Grid computing context places thiss research in the arena of the Semantic Grid [2].

2 The eBank UK Project - outcomes to date

The eBank UK project is addressing this challenge by investigating the role of aggregator services in linking data-sets from Grid-enabled projects to open data archives contained in digital repositories through to peer-reviewed articles as resources in portals. This innovative JISC-funded project which is led by UKOLN in partnership with the Universities of Southampton and Manchester, is seeking to build the links between e-research data, scholarly communication and other on-line sources. It is working in the chemistry domain with the EPSRC funded eScience testbed CombeChem [3], which is a pilot project that seeks to apply the Grid philosophy to integrate existing structure and property data sources into an information and knowledge environment.

The specific examplar chosen from this subject area is that of crystallography as it has a strict workflow and produces data that is rigidly formatted to an internationally accepted standard. The EPSRC National Crystallography Service (NCS) is housed in the School of Chemistry, University of Southampton, and is an ideal case study due to its high sample throughput, state of the art instrumentation, expert personnel and profile in the academic chemistry community. Moreover, recent advances in crystallographic technology and computational resources have caused an explosion of crystallographic data, as shown by the recent exponential growth of the Crystal Structure Database (CSD) see Cambridge Crystallographic Data Centre. However, despite this rise it is commonly recognized that approximately only 20reaching the public domain. This situation is even worse in the high throughput NCS scenario where approximately 15disseminated, despite producing ¿60 peer reviewed journal articles per

annum. With the imminent advent of the eScience environment, this problem can only get more severe.

A schema for the crystallographic experiment has been devised that details crystallographic metadata items and is built on a generic schema for scientific experiments, figure 2. During the deposition of data in a Crystallographic EPrint metadata items are seamlessly extracted and indexed for searching at the local archive level. The top level document includes 'Dublin Core bibliographic' and 'chemical identifier' metadata elements in an Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) compliant form, which allow access to a secondary level of searchable crystallographic metadata items, which are in turn directly linked to the associated archived data. In this manner the output from a crystallographic experiment may be disseminated as 'data' in such a way that aggregator services and researchers may add value to it and transform it into knowledge and the publication 'bottleneck' problem can be addressed.

The metadata about datasets made available in the repository will be harvested into a central database using the OAI-PMH. This metadata will then be indexed together with any other available metadata about research publications. A searchable interface will enable users to discover datasets as well as related literature. The metadata contains links back to the datasets which the user will be able to follow in order to obtain access to the original data, when this is available. Harvested metadata from different repositories not only provides a common entry point to potentially disparate resources (such as datasets in dataset repositories and published literature which may reside elsewhere) but also offers the potential of enhancement of the metadata such as the addition of subject keywords to research datasets based on the knowledge of subject classification terms assigned to related publications. A further area of work investigates the embedding of the search interface within web sites, adopting their look-and-feel. PSIgate (http://www.psigate.ac.uk/) will be used to pilot these embedding techniques based on CGI-mechanisms and portal-related standards.

The concept we have implemented within the Southampton e-print archive system is Data Publication@Source [4]. Crystallographic EPrints use the OAI concept to make available ALL the data generated during the course of a structure determination experiment. That is the publishable output is constructed from all the raw, results and derived data that is generated during the course of the experiment. This presents the data in a searchable and hierarchical system that relates to the workflow of the experiment. This metadata includes bibliographic and chemical identifier items which are above a secondary level of searchable crystallographic items which are directly linked to the associated archived data. The table below depicts the schema and shows the hierarchical manner in which the open archive report is constructed in the figure below.

Hence the results of a crystal structure determination may be disseminated in a manner that anyone wishing to utilise the information may access the entire archive of data related to it and assess its validity and worth. In the future a notification, or bulletin board type system, (much like Amazons reviewers comments) could be added to create in effect a new type of peer reviewed publication.

Name	Files associated with this stage			Metadata associated with this stage	
	File	Type	Description	Name	Data Type
Initialisation	.htm	HTML	Metadata for crystallography expt	Morphology	*STRING (SET)
	i*.kcd	BINARY	Unit cell determination images	Solvent	*STRING
			Unit cell	Sample_image	.JPG
Collection	s*.kcd	BINARY	Diffraction images	Instrument_Type	*STRING
	scan.jpg	JPG	Visual version of .kcd file	Temperature	*INTEGER
				Crystal_image	.JPG
				Software_Name	STRING
				Software_Version	INTEGER
Processing	scale_all.in	ASCII	Result of processing	Cell_a	*NUMBER
	scale_all.out	ASCII	Result of correction on processed data	Cell_b	*NUMBER
	.hkl	ASCII	Derived data set	Cell_c	*NUMBER
	.htm	HTML	Report file	Cell_alpha	*NUMBER
				Cell_beta	*NUMBER
				Cell_gamma	*NUMBER
				Crystal_system	*STRING (SET)
				Completeness	*INTEGER (%)
				Software_Name	STRING
				Software_Version	INTEGER
Solution	.prp	ASCII	Symmetry file, log of process	Space_group	*STRING (SET)
	xs.lst	ASCII	Solution log file	Figure_of_merit	*NUMBER
				Software_Name	STRING
				Software_Version	INTEGER
Refinement	xl.lst	ASCII	Final refinement listing	R1_obs	*NUMBER
	.res	ASCII	Output coordinates	wR2_obs	*NUMBER
				R1_all	*NUMBER
				wR2_all	*NUMBER
				Software_Name	STRING
				Software_Version	INTEGER
CIF	.cif	ASCII	Final results	Formula_moiety	*STRING
	checkcif.htm	HTML	Automatic validation results	CIF_check	*STRING
Report	.html	.HTML	Publication format	EPrint_type	*CRYSTAL
			(HTML/XHTML)		STRUCTURE
				Authors	*STRING
				Affiliations	*STRING
				Formula_empirical	*STRING
				Compound_name	*STRING
				CCDC_Code	*STRING
				Compound_class	*STRING (SET)
				Keywords	*STRING (SET)
				Available_data	*STRING (SET)
				Related_publications	STRING

Fig. 2. The draft version of the schema details for the crystallographic data showing the main data types and associated file names generated in a crystal structure determination.

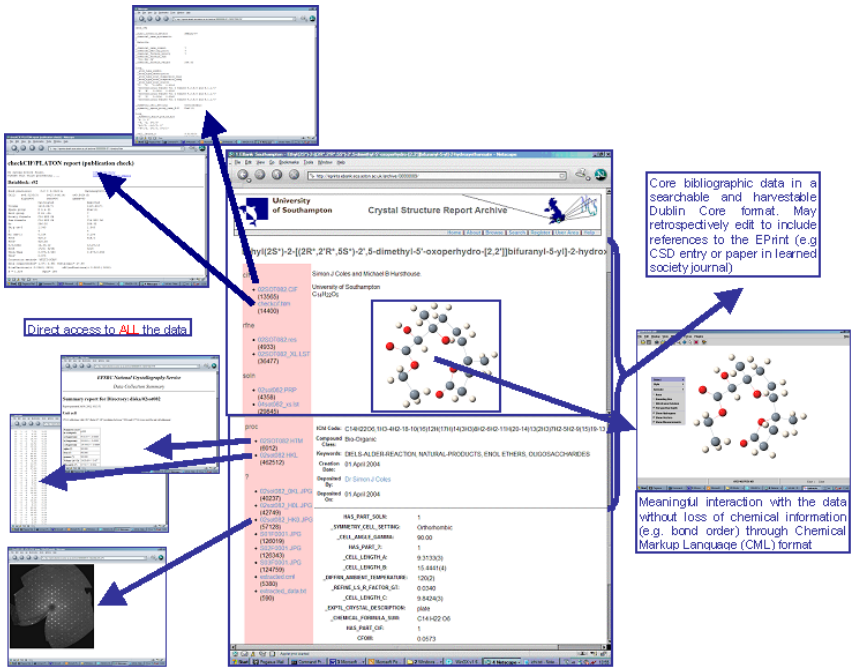

Fig. 3. A summary of the different screens available from the basic e-data report showing how the data and metadata described in the basic schema of figure 2 are displayed.

3 Conclusions - issues and challenges for the future

While only test versions of the e-crystal-data-report have been produced so far they have been populated with real crystallographic data sets produced at the National Crystallography Service (NCS). The lessons learnt so far fall in to three main areas. First, the ease of use in inputting the data to the system and ensuring that sufficient information about the materials, the experiment and the nature of the data is being entered. Some of the input styling is taken over directly from the current e-print system which of course allows for files of different types to be up loaded. However the issue of file types becomes more complex with the advent of data. In this regard the conformity of the crystallography community is an advantage, with common file types with well-understood extensions. Never the less some standardization is imposed following the NCS practice at this stage. Conversion between different file types for the higher-level data elements (e.g. the molecular structure files) can be done automatically, with a high degree of fidelity, and this is necessary to provide suitable data for the visualization programs built in to the viewing interface.

This brings us on to the second main area, the viewing of this information directly from the e-print system. Once an entry is located then the information is presented on a web interface with the major details about the molecule available, including a visual, rotateable image. The conversion of the crystallographic CIF file to other formats (e.g. MOL file) is necessary for this. Similarly the newly define unique chemical identifier (INCHI) is also calculated by converting the CIF file via these stages to a CML file. The required conversions are implemented (or soon will be) as web services to allow updates independently of the main e-print software system.

The final issues revolve around finding the existence of the data on the e-print system. As indicated above the data schema that provides the outline for the arrangement of the data highlights the significant data that a chemist would wish to search on. This is made available via an OIA interface to harvesting programs, which can then provide additional functionality to enable a multi-parameter search. This side of the system is less tested and does present some problems for us as many of the current search systems allow for searches on chemical structures (drawn in 2 or 3D) using proprietary algorithms, especially for the sub-structure search. We are thus not able to implement this type of search but can demonstrate that the data needed for such searches can be made available to data aggregators.

The test experiments have proved very successful and have engaged the interest of several of the traditional suppliers of crystallographic information, who wish in effect to move up market in the information supply chain. In the academic community there is a growing word wide interest. The system of separating the archiving of the data, including all the raw and background information, from the aggregated higher level structural data which is to be curated over a much longer term, looks to be very successful.

We should stress that the e-experiment-data-reports are not restricted to crystallographic data. These data sets were chosen as an ideal test case due to

their relative uniformity, community agreement and availability of diagnostics to enable the user to assess the quality of the data. We are now extending the system to cover spectroscopic data, for which again the issue of the lack of extensive libraries of even common molecules is a major hindrance to efficient research. We are applying the principle to Raman spectra are working with one of the spectrometer supplies from the start to ensure that the links to the e-print system can be built in to the spectrometer software, further simplifying the task of disseminating the data.

One aspect that may need further consideration if the automated processes are enhanced, is finer control over access to the data by different groups at different times. This will of course overlap with security concerns.

References

1. Lyon, L.: Developing Information Architectures - to support Research, Learning and Teaching. In: UCISA Conference. (2002)
2. De Roure, D., Jennings, N.R., Shadbolt, N.R.: The Semantic Grid: A future e-Science infrastructure. In: Grid Computing - Making the Global Infrastructure a Reality. John Wiley and Sons Ltd (2003) 437–470
3. Frey, J.G., Bradley, M., Essex, J.W., Hursthouse, M.B., Lewis, S.M., Luck, M.M., Moreau, L., De Roure, D.C., Surridge, M., Welsh, A.: Combinatorial chemistry and the Grid. In: Grid Computing - Making the Global Infrastructure a Reality. John Wiley and Sons Ltd (2003) 945–962
4. Frey, J., De Roure, D., Carr, L.: Publishing at Source: Scientific Publication from a Web to a Data Grid. In: EuroWeb 2002 Conference, Oxford (2002)
5. Bearman, D., Lytle, R.: The Power of the Principle of Provenance. Archivaria (**21**) 14–27
6. Hitchcock, S., Brody, T., Gutteridge, C., Carr, L., Hall, W., Harnad, S., Bergmark, D., Lagoze, C.: Open Citation Linking: The Way Forward. D-Lib Magazine **8** (2002)