Integrating research data into the publication workflow: the eBank UK experience

Rachel Heery, Monica Duke, Michael Day, Liz Lyon[(1)], Michael B. Hursthouse, Jeremy G. Frey, Simon J. Coles[(2)], Christopher Gutteridge, Leslie A. Carr [(3)]

[(1)]UKOLN
University of Bath, BA2 7AY, United Kingdom
Email: r.heery, m.duke, m.day, l.lyon@ukoln.ac.uk

[(2)]School of Chemistry,
University of Southampton, Southampton SO17 1BJ, United Kingdom
Email: m.b.hursthouse ,j.g.frey, s.j.coles@soton.ac.uk

[(3)]School of Electronics and Computer Science,
University of Southampton, Southampton SO17 1BJ, United Kingdom
Email: cjg, lac@ecs.soton.ac.uk

INTRODUCTION

E-science has the potential to enrich scholarly communication from the perspective of both research and learning. This paper will explore emerging infrastructure that will enable effective curation of data to deliver this potential. The exponential growth in digital data arising from e-science requires new modes of data curation. The large volume of digital data now being created calls for new modes of publication, discovery, re-use and preservation. New patterns of **publication**, 'publication at source', are required, embedding the publication of data into the research scientist's work patterns. Curation of digital data must support **discovery** services, building new service models on traditional library and publishing practice. There are opportunities for **re-use** of the increasing amount of digital data being produced, whether re-use takes place as part of research activity or in the development of course materials. In addition, secure archiving is necessary to ensure **preservation** of digital data.

The re-use and sharing of original data is a fundamental tenet of e-science. Successive levels of derived data will be produced from original data through refinement, transformation, interpretation, and generalisation. Original data is beginning to be made available in publicly available databases, whether institutional repositories, or publishers' databases. This is happening particularly in certain specialisms, for example in the bio-informatics field, protein and genome sequences databases, and in chemistry, material safety information, and, significantly for eBank UK, crystal structures.

As new methods and technologies emerge to support the research data life-cycle through publication, discovery, re-use and preservation, there is a need to manage original and derived data (datasets produced both by experimentation and data re-use) alongside bibliographic data in an integrated fashion. Technologies underlying the emerging global network infrastructure (Internet, World Wide Web, Semantic Web, Grid) increasingly will support links between the activities of e-research and e-science, digital libraries and e-learning, offering enhancements to scholarly communications throughout the life-cycle from experimental research to learning. There is increasing interest in establishing institutional repositories to provide services to manage the variety of digital materials that form the intellectual output of educational and research institutions [1]. Significantly such institutional repositories also have a central role in taking forward the open access agenda, supporting e-print archives that enable self-archiving of journal article pre-prints and post-prints, and potentially other categories of material, by academic staff [2].

This changing landscape offers opportunities to curate scientific datasets more effectively, improving access and integration between services. The eBank UK project is addressing this challenge by investigating the role of aggregator services in linking metadata describing e-prints of peer reviewed journal articles to datasets from Grid-enabled projects made available within institutional and publisher e-data repositories.

THE EBANK UK PROJECT

The Joint Information Systems Committee (JISC) funded eBank UK project is led by UKOLN in partnership with the Universities of Southampton and Manchester. The project is working in the chemistry domain with the EPSRC (Engineering and Physical Sciences Research Council) funded e-science test-bed Combechem [3] at the University of Southampton, a pilot project that seeks to integrate existing structure and property data sources into an information and knowledge environment. The eBank UK demonstrator is being developed within a particular chemistry domain, crystallography, with a view to assessing, in the longer term, the feasibility of a generic approach across other disciplines.

The Combechem project offers an ideal research test-bed as it generates large quantities of digital data including crystallography data and physical chemistry data. Within chemistry research the vast majority of measurements in experiments are obtained using computer controlled instruments, so both data and metadata is provided automatically in digital form. Such data must be properly curated so a well-managed archive is vital, an archive that will outlast the completion of research projects and possible changes in staffing. Establishing institutional repositories for scientific data, maintained by a university or research institute, offers a possible solution. In order to populate such institutional e-data repositories effectively, the creation and archiving of both data and metadata must be fully integrated into the experimental workflow. Combechem is a test-bed for this approach, with a focus on crystallography and spectroscopic data as exemplars. eBank UK has chosen to focus on crystallography as this area of research has a strict workflow and produces data that is rigidly formatted to an internationally accepted standard, namely the Crystallographic Information File (CIF) standard [4], maintained by the International Union of Crystallography [5].

EBANK UK MOTIVATION

At present, as a result of technological advances in instrumentation and with the advent of e-science and high throughput philosophies, a publication bottleneck exists in many scientific communities. Accordingly only a small percentage of the data generated by many scientific experiments appears in, or is referenced by, the published literature. In addition, publication in the mainstream literature still offers only *indirect* (and often expensive) access to this data. As a consequence the user community is deprived of valuable information and funding bodies are getting a poor return for their investments. By moving towards a ‘publication at source’ approach access and discovery of research data would be made easier, which in turn would encourage re-use of data in further experiments.

Currently once a research experiment is finished the initial dissemination may be via a letter or communication, followed later by a more detailed explanation in a full paper. eBank UK will demonstrate how data might be ‘published at source’ directly in open archives and subsequently linked to either peer- reviewed journal articles or automatically deposited entries in specialised databases. Such an approach would enable the ‘publication’ of the enormous amount of data that is being generated within research areas such as crystallography. Publishers of crystal structures, and indeed most types of scientific data, make available to subscribers limited amounts of results data, such as CIFs. However this is only offering a partial solution, as the final result in most cases is merely a fraction of the digital data generated during the course of the experiment. Moreover this single result does not allow the 'reader' to assess the interpretations and assumptions made by the experimenter during the data workup. This is due to traditional publishing protocols being time consuming and considerably delaying, incapable of keeping up with the current data explosion. In addition such services are selective, and do not make available all the metadata that is associated with the primary data.

In parallel to increasing interest in archiving data, there is growing policy support and funding to encourage researchers to self-archive peer-reviewed journal articles in institutional repositories. This offers an opportunity to integrate services offered by e-data archives and journal article e-print archives. As the availability of journal article e-prints increases, users would benefit from a service linking back from the article to the raw or processed data. Various services might be built based on metadata made available through institutional repositories: providing context sensitive linking within electronic versions of journal, links from e-print archives to datasets, linking to datasets from e-learning portals.

The eBank UK demonstrator shows how an e-prints repository of peer-reviewed journal articles could include links to associated research data. In particular eBank UK focuses on the potential role for aggregator services that might harvest metadata describing datasets from institutional repositories as contributed by Grid-enabled projects.

EBANK UK ARCHITECTURE

The eBank UK project is exploring linking from metadata describing journal articles to the original datasets, thereby integrating access to datasets from Open Access e-print services. The aim of the project is to develop a pilot service. This will be done by

- Creating an institutional repository of crystallography data, an open e-data repository (at the University of Southampton)
- Modifying e-prints.org software [6] both to support the open e-data archive, and to enable harvesting of metadata describing datasets (software developed and maintained by the University of Southampton)
- Demonstrating an eBank UK search service linked to ePrints UK [7], indexing harvested descriptions of datasets and Journal articles (at UKOLN, University of Bath)
- Embedding the eBank UK service into the PSIgate [8] subject gateway (at the University of Manchester)

The architecture of eBank UK adopts the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [9] which offers a specific design approach. OAI-PMH envisages a class of 'data providers' acting as repositories of resources. These data providers are accessed programmatically by means of a number of requests (which together make up the OAI-PMH). A second class of 'service providers' issues the OAI-PMH requests to the data providers. A series of requests made by the service provider (the full details of which are beyond the scope of this paper) culminates in 'harvesting' of the metadata about the resources held by the data provider. In other words, resource descriptions held by the data provider are sent in response to the service provider requests. By selectively making requests to one or more data providers, the service provider aggregates a collection of resource descriptions. The service provider (or aggregator) is thus positioned to provide a single point of entry to disparate repositories of resources and resource descriptions, and can offer a variety of services to support the discovery of published data and literature.

The harvesting design addresses some well-known limitations of cross-searching techniques. Cross-searching techniques generally send specific search requests in parallel to different sources (by some specified protocol) and combine the various responses into a result for the cross-search. In contrast, search services built on harvested metadata carry out local searches on the pre-harvested metadata. This circumvents two problems which may be encountered in cross-searching: firstly delayed responses necessitate that the requester decides whether to wait on all responses before presenting any results (for example before applying any ranking) and secondly the potential unavailability of the search target at the moment the request is made may leave a gap in the response, corresponding to the unavailable target.

For historical reasons, the OAI-PMH is very closely tied with the Open Access philosophy of publication, whereby the dissemination of publications (such as journal articles) is improved by depositing the item in an institutional or subject-based repository. The repository acts as a central focus from which the article, together with the associated metadata, may be shared freely with the research community. The very first Open Access collections (e.g., arXiv [10], CogPrints [11]) provided the test-bed on which the OAI-PMH was developed and refined, until it evolved into the current stable version. Partly as a consequence of this association, OAI-PMH-compliant repositories have proliferated, and a number of Open Access collections fulfilling the role of data providers (in an OAI-PMH sense) are currently available. These data providers are distributed in the US, Europe, Australia and wider. They take on a different focus according to their context; many are institutional and are seen as a central point of dissemination for an institution's published assets; however different flavours of repositories are possible, for example centred on images, theses and even CVs.

Compared to the number of data providers, the emergence of service providers has been slower. Arguably this is an unavoidable situation. A necessary level of data provider availability is required to provide the material to fulfil service provider requests. A range of data provider instances would also be needed to motivate the development of different service provider models. A lag-time may be inevitable and service provider maturity may follow on the success of the data provider take-up.

As stated, service providers act as 'aggregators', selectively harvesting metadata from one or more data repositories. The OAI-PMH enables a degree of selective harvesting based on a number of factors. At the most basic (and mandatory) level, OAI-PMH supports requests based on the last-updated date, so that cumulative harvesting of metadata can be performed. Resources can also be optionally partitioned into 'sets', defined according to a manner appropriate to the data provider (e.g. by subject or by resource type). Finally, service providers can be selective in deciding which data providers to harvest, for example only harvesting from collections that meet quality criteria by enforcing policies about the material deposited. Alternatively a service provider could concentrate on meeting the

needs of a specific community with an interest in certain kinds of resources. Furthermore, the service provider can tailor the searching services to the searching and discovery requirements of the particular user group that it serves. In this manner, service provides can differentiate the services on offer.

The ePrints UK service, funded by JISC, is providing a focus for the cross-searching of assets made available by institutions, presenting a subject-based service targeting the existing user base of JISC-funded subject-based resource discovery services. Furthermore, e-prints UK is seeking to enhance the harvested metadata through citation analysis, name authority services and automated subject classification.

In the eBank UK project, this Open Access philosophy is being extended to research data. A data repository has been created at the University of Southampton. Crystallographers deposit the datasets produced at the National Crystallography Service in an enhanced version of the e-prints.org software that accommodates self-archiving by chemists. During the deposit process, metadata about the datasets is entered or generated automatically. This metadata is then made available in the repository that, in addition to supporting local searching, browsing and downloading of datasets, also acts as an OAI-PMH compliant data provider. eBank UK has also implemented an example of a service provider that harvests metadata about datasets from the repository (using OAI-PMH). This metadata provides the basis on which to create innovative services that support the discovery and re-use of datasets, as well as creating links with the published literature. The shared infrastructure re-using the OAI-PMH supports interoperability between systems disseminating publications and those disseminating research datasets.

EBANK UK DATA FLOW

Fig. 1 summarises the flow of data and metadata within the eBank UK project. Datasets are created during experimental processes and are instantiated as a number of data files. The files are self-archived in an OAI-PMH repository and descriptive metadata is created as part of the deposit process. The metadata consists of a number of fields that are either entered either by the depositor or generated automatically by the modified e-prints.org software. In the institutional eData repository, a local web interface (using HTML) is presented to the user. This interface consists of crystallographic eData reports, an interactive visualization of a derived crystal structure, selected values that describe the experiment and the results, as well as links to all the data generated in the course of an experiment.

The eBank UK aggregator harvests the metadata from this data repository by making OAI-PMH requests. The metadata schema used for exchange is discussed in the next section. An SGML indexing and searching engine, Cheshire [12], is then used to provide indexing and searching functionality over the metadata. When combined with

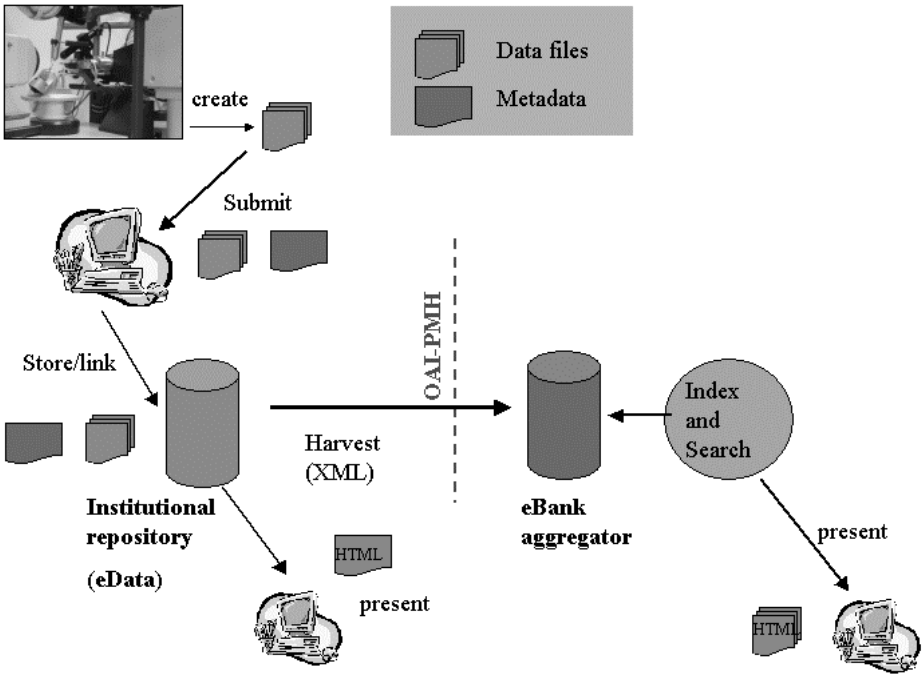

Fig. 1. Data flow in the eBank UK architecture

publication metadata, any links made between different datasets, or datasets and publication can be revealed to the user in the display of the search results. Links can be made indirectly, for example by identifying common occurrences of author names, keywords or other subject-specific terms (such as chemical identifiers in our particular crystallography exemplar). Alternatively, direct links can be created if there is a specific reference, in the dataset or the publication metadata, to the related work. For example, an OpenURL could be generated to link to an instance of a journal article that is available electronically in a user's institution, if sufficient metadata about this related article is included in the metadata record describing the data set.

One further area of work in the eBank UK project is the embedding of the search service into external services to reach the user base by means of alternative points of entry. There are a number of mechanisms by which this can be achieved. All the mechanisms would allow a third party service to make requests to the service provider, which in turn returns results in a manner that can be manipulated by the third party to fit the look and feel of their interface, for example the University portal or a subject portal. EBank UK has implementation experience of using CGI-mechanisms, web service protocols (SRW Search/Retrieve Web Service [13]) and portal technologies. These techniques will be used to present the search service through PSIgate, which has an audience in higher and further education, mainly in the UK, but also worldwide, with an interest in the physical science. The eBank UK search service will complement the other features of PSIgate, adding discovery and access of research data to the landscape of scientific services already on offer at PSIgate.

EBANK SCHEMAS

The services that can be built on aggregated metadata can only be as good as the metadata that is available to them. To achieve interoperability between dataset and publication metadata, there must exist some consensus on the data models and the metadata schemas being used to exchange metadata. Furthermore, to support discovery facilities between datasets from different communities, there must be some agreed commonality in the models and schemas if cross-disciplinary services are to be built.

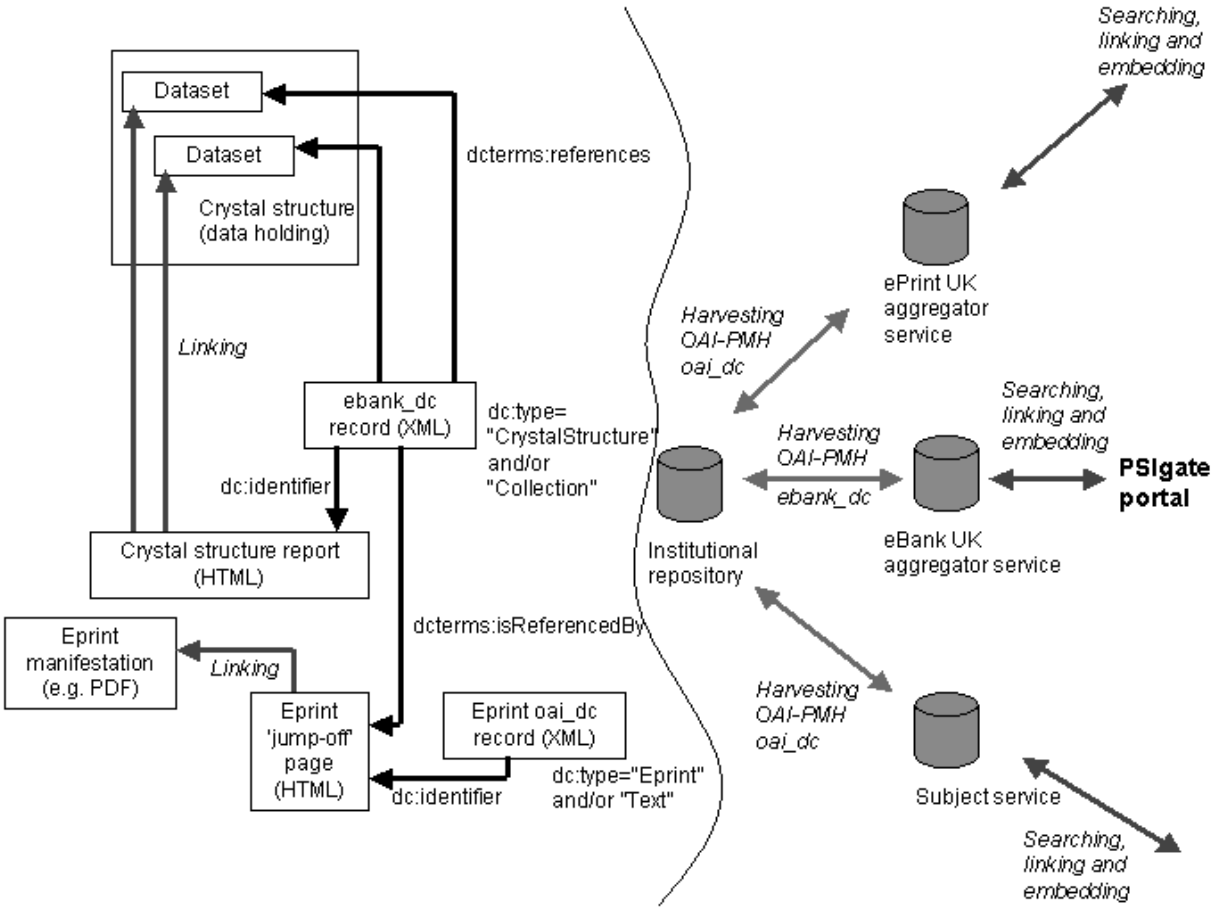

Fig. 2 . The links between metadata records, HTML views of the records and data sets (left) and the eBank UK architecture (right)

As a starting point the eBank UK project has taken the Dublin Core metadata element set [14], and other DC terms available [15], and adapted them to describe the datasets that are deposited by crystallographers in the institutional eData repository. The reasons for this are two-fold. Firstly, simple DC is mandated within the OAI-PMH (although the protocol also encourages the exposure of specialized metadata records alongside the minimally required DC). Secondly, DC was conceived to provide a minimum baseline of interoperability, which can be extended and specialised for specific needs.

The metadata schema uses elements from the fifteen elements of the Dublin Core Element Set as the basis for descriptions. For example the creators of the datasets are designated within the Dublin Core creator element. Discussion with the users revealed that there are a number of different ways to describe the subject of the datasets. Crystallography experiments revolve around a single molecule, which can be thought of as the 'topic' of the experiments. There are a number of established ways of identifying molecules, which include internationally recognised methods of specifying their formulas or names. These different vocabularies have been incorporated into the schema through the encoding schemes facility of qualified Dublin Core. The Dublin Core Metadata Initiative (DCMI) provides recommendations for including terms from vocabularies in encoded XML suggesting "*Encoding schemes* should be implemented using the 'xsi:type' attribute of the XML element for the *property*." An example from the eBank UK metadata records for datasets is the representation of the chemical empirical formula as:

<dc:subject xsi:type="eBank UKterms:empiricalFormula">C288 H200 Cl24 F48 O48 P16 Pd16</dc:subject>

Efforts are ongoing in the chemistry community to agree on namespaces for these vocabularies [16]. As these namespaces emerge, they can substitute the eBank UK namespace designations, which have been used temporarily until the standard recommendations become available.

The metadata record harvested by the eBank UK aggregator describes the collection of datasets which are identified by a URL that links to the HTML entry page in the data repository where they were deposited. This identifier is used as the value of the DC identifier element. The location for individual datasets can also be given in DC relation elements. Since it is desirable to be able to describe the type of datasets, the extended DC vocabulary mechanism was once again used to specify types.

<dc:type xsi:type="eBank UKterms:EBank UKDatasetType">CollectionDataset</dc:type>

Related literature, such as journal articles, can also be exposed in the metadata record, either by citing a unique identifier (e.g. DOI) or by textual citation (author details, title, issue etc.).

However this metadata schema (and the Dublin Core model) does not accommodate the whole complexity of the model of datasets. There may be some inherent relationships between datasets that would be usefully revealed in the metadata. For example in the case of crystallography, datasets are related to one another by sequence since they are generated (by

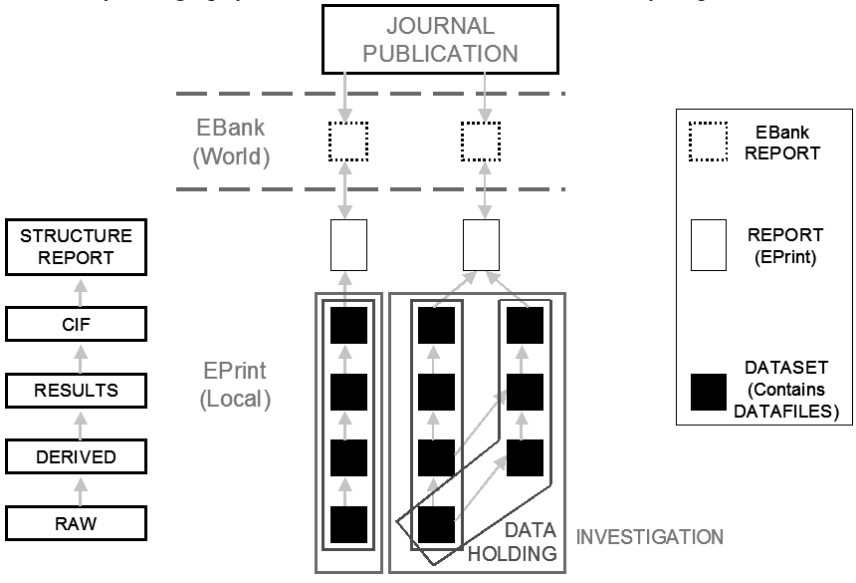

Fig. 3. The crystallography work flow and relationship between data sets

measurement or analysis) from a series of consequent stages in the experimental process. Similarly, datasets may be made up of one or more files, which may again be related in specific ways, or crucially, may be stored in different locations or facilities, governed by varying access control mechanisms.

The CCLRC data model [17] developed by the Council for the Central Laboratory of the Research Councils, is an emergent example that attempts to deal with this complexity by describing the relationship between experiments, investigators, data holdings, datasets, data files, logical and physical locations. Another area of interest is provided by recently proposed schemas for describing complex objects, such as the Metadata Encoding and Transmission Standard (METS), the MPEG21 Digital Item Declaration Language (DIDL) [18] and content packaging standards from the e-learning community.

CONCLUSION

There are increasing commonalities in the concerns of those developing digital libraries and those creating and managing e-science data. This includes the potential use of common technologies for managing datasets and bibliographic data e.g. the OAI-PMH; and overlaps between services that might be offered based on the data e.g. linking datasets and journal articles

During its initial phase, eBank UK has focused exclusively on the chemistry domain and in particular on crystallography. There is potential to expand its remit to a wider range of crystallography data, to other chemistry sub-domains and indeed to encompass the broader range of physical sciences. A longer term ambition would be for other scientific communities, including social sciences and humanities, to adopt a similar approach. eData repositories might be supported, both by institutions (universities and research institutes) and publishers. In order to progress this vision an immediate challenge for the wider community is to reach consensus on a common data model for scientific datasets.

Agreement would be required on a common approach to the range of services that might be built on such repositories. Value added services might include the enhancement of data with visualisation and scientific context through the use of mark up languages (e.g., Chemical Markup Language (CML) [19, 20], Computational Chemistry Markup Language (CCML), Mathematical Markup Language (MathML) [21]). Other services more typically associated with digital libraries might be offered, such as enhancement of metadata by automated subject classification and authority control. The provision of services connected with subject access such as these, would be particularly appropriate in the context of subject portals, and eBank UK is indeed exploring this area in collaboration with the Resource Discovery Network PSIgate service.

There are other outcomes resulting from enhancements to curation of scientific data that are outside the scope of this paper. These include the pedagogical benefits of providing access to primary research data within eLearning materials. Within the context of eBank UK, this would be possible within postgraduate courses in chemistry, undergraduate project work, or chemical informatics courses. In a wider context there could be inclusion of e-research data in e-learning courses, through links in reading lists, essay assignments, analytical problem solving, and through practical work. Enhanced availability of and access to original and derived scientific data might also suggest possible changes in the peer review process. The provision of all associated data in a repository might satisfy most scientists that the formal refereeing of data via peer review could be removed from the publication chain, since the data can be assessed readily on line by anyone at any time.

In order to move from demonstrator to service, there is a need to involve publishers and other existing specialist services, as well as institutions, in the eBank UK approach. There is potential for a variety of business models whereby publishers might build services based on harvested metadata, particularly those publishers that already have a focus on access to original data.

In conclusion, the eBank UK project has built on a joint approach arising from both the digital library community, the Grid and computer science community. eBank UK demonstrates benefits to the research community, and has shown the potential for integration into digital library services. For more information on the eBank UK project, please see the project web pages at http://www.ukoln.ac.uk/projects/ebank-uk.

ACKNOWLEDGEMENT

The eBank UK project is funded by JISC under the Semantic Grid and Autonomic Computing Programme.

REFERENCES

[1] C. Lynch, "Institutional repositories: essential infrastructure for scholarship in the digital age," *ARL Bimonthly Report*, no. 226, February 2003. Retrieved July 22, 2004, from: http://www.arl.org/newsltr/226/ir.html

[2] R. Crow, *The case for institutional repositories: a SPARC position paper*. Washington, D.C.: Scholarly Publishing & Academic Resources Coalition, 2002. Retrieved July 22, 2004, from: http://www.arl.org/sparc/IR/ir.html

[3] J.G. Frey, M. Bradley, J.W. Essex, M.B. Hursthouse, S.M. Lewis, M.M. Luck, *et al*., "Combinatorial chemistry and the Grid," in *Grid computing: making the global infrastructure a reality*. F. Berman, G. Fox and A.J.G. Hey, Eds. Chichester: Wiley, pp. 945-962, 2003.

[4] S.R. Hall, F.H. Allen and I.D. Brown, "The Crystallographic Information File: a new standard archive file for crystallography," *Acta Cryst*., vol. A47, pp. 655-685, 1991.

[5] International Union of Crystallography. Retrieved July 22, 2004, from: http://www.iucr.org/

[6] The eprints.org software. Retrieved July 22, 2004, from: http://software.eprints.org/

[7] ePrints UK project. Retrieved July 22, 2004, from: http://www.rdn.ac.uk/projects/eprints-uk/

[8] The PSIgate service Retrieved July 22, 2004, from http://www.psigate.ac.uk/

[9] C. Lagoze, H. Van de Sompel, M. Nelson and S. Warner, Eds., *Open Archives Initiative Protocol for Metadata Harvesting*, version 2.0, June 2002. Retrieved July 22, 2004, from: http://www.openarchives.org/

[10] arXiv.org e-Print archive. Retrieved July 22, 2004, from: http://arxiv.org/

[11] CogPrints. Retrieved July 22, 2004, from: http://cogprints.ecs.soton.ac.uk/

[12] Cheshire SGML search engine. Retrieved July 22, 2004, from: http://cheshire.berkeley.edu/

[13] SRW Search/Retrieve Web Service. Retrieved July 22, 2004, from: http://www.loc.gov/z3950/agency/zing/srw/

[14] Dublin Core Metadata Initiative, *Dublin Core Metadata Element Set*, version 1.1, DCMI Recommendation, June 2003. Retrieved July 22, 2004, from: http://dublincore.org/documents/dces/

[15] Dublin Core Metadata Initiative, *DCMI Metadata Terms*, DCMI Recommendation, June 2004. Retrieved July 22, 2004, from: http://dublincore.org/documents/dcmi-terms/

[16] XML Data Dictionaries in Chemistry. Retrieved July 22, 2004, from http://www.iupac.org/projects/2002/2002-022-1-024.html

[17] S. Sufi, B. Matthews and K. Kleese van Dam, "An interdisciplinary model for the representation of scientific studies and associated data holdings," UK e-Science All Hands Meeting, Nottingham, UK, 2-4 September 2003. Retrieved July 22, 2004, from: http://www.nesc.ac.uk/events/ahm2003/AHMCD/pdf/020.pdf

[18] J. Bekaert, P. Hochstenbach and H. Van de Sompel, "Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Library Digital Library," *D-Lib Magazine*, vol. 9, no. 11, November 2003. Retrieved July 22, 2004, from: http://www.dlib.org/dlib/november03/bekaert/11bekaert.html

[19] P. Murray-Rust and H.S. Rzepa, "Chemical markup, XML, and the Worldwide Web, 1. basic principles," *J. Chem. Inf. Comput. Sci.*, vol. 39, pp. 928-942, 1999.

[20] P. Murray-Rust and H.S. Rzepa, "Chemical markup, XML, and the World Wide Web, 4. CML Schema," *J. Chem. Inf. Comput. Sci.*, vol. 43, pp. 757-772, 2003.

[21] D. Carlisle, P. Ion, R. Miner and N. Poppelier, Eds., *Mathematical Markup Language (MathML) Version 2.0*, 2nd ed. W3C Recommendation, October 2003. Retrieved July 22, 2004, from: http://www.w3.org/TR/MathML2/