Enhancing Access to Research Data : the Challenge of Crystallography

Monica Duke, Michael Day, Rachel Heery
UKOLN
University of Bath
Bath BA2 7AY, UK
{m.duke, m.day, r.heery}@ukoln.ac.uk

Leslie A. Carr
School of Electronics and Computer Science
University of Southampton
Southampton SO17 1BJ, UK
lac@ecs.soton.ac.uk

Simon J. Coles
School of Chemistry
University of Southampton
Southampton SO17 1BJ, UK
s.j.coles@soton.ac.uk

ABSTRACT
This paper describes an ongoing collaborative effort across digital library and scientific communities in the UK to improve access to research data. A prototype demonstrator service supporting the discovery and retrieval of detailed results of crystallography experiments has been deployed within an Open Archives digital library service model. Early challenges include the understanding of requirements in this specialized area of chemistry and reaching consensus on the design of a metadata model and schema. Future plans encompass the exploration of commonality and overlap with other schemas and across disciplines, working with publishers to develop mutually beneficial service models, and investigation of the pedagogical benefits. The potential improved access to experimental data to enrich scholarly communication from the perspective of both research and learning provides the driving force to continue exploring these issues.

Categories and Subject Descriptors
H.3.7 [Digital Libraries] Collection, Dissemination, Standards, User issues

General Terms
Experimentation, Standardization

Keywords
Eprints.org, crystallography, metadata, Dublin Core, OAI-PMH, scholarly communication, institutional repositories.

1. INTRODUCTION
Modern e-science produces increasingly large volumes of data as computational tools enable experiments to be performed more frequently and more efficiently. In crystallography in the 1960s, a doctoral student might have investigated three or so structures, now this number can be analyzed in a single morning, yet the publishing protocols for reporting this work are essentially unchanged. Across the scientific domain, only a small percentage of data generated by many scientific experiments appears in, or is referenced by, the published literature [33]. 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 get a poor return on investments.

The underlying motivation for eBank UK is to demonstrate 'publication at source', the rapid dissemination of structural information to the scientific community by means of new modes of service provision. Typically journal publication has been detached from the production of the experimental data, with the result that managing and providing access to full experimental data has not been simple. A journal article describing the results of scientific work is typically a distillation of experimental data aimed at a wider audience than the immediate peers of the authors. The article will often be concerned with placing experimental work in its context and will reduce reporting of the data to the most significant results, often expressed in reduced graphical or tabular form. Immediate peers in the discipline, however, may require access the original data to verify reproducibility or to build on those data. Although some journals have attempted to store data relating to published articles, typically this data is only a partial set of the complete dataset. Many journals, especially those based on print formats, do not have the space for storing large sets of data. For the research chemist, just 300,000 crystal structures are available in subject specific databases that have harvested their content from the published literature. It is estimated that 1.5 million structures have been determined in research laboratories worldwide and hence less than 20% of data generated in crystallographic work is reaching the public domain [5]. This shortfall is entirely due to current publication mechanisms. As high-throughput technologies, automation and e-science become embedded in chemical and crystallographic working routines, the publication bottleneck can only become more severe [34].

The Joint Information Systems Committee (JISC) funded the eBank UK project [20], a joint effort between crystallographers, computer scientists and digital library researchers, approached this problem area by investigating the contribution existing digital library technologies could make. Experimental scientific data are produced electronically, so are immediately amenable to digital storage, aggregation and discovery – broadly speaking to 'digital curation'. Within the digital library community institutional repositories are emerging as a focus for the curation of the variety of digital materials that form the intellectual output of educational and research institutions. The eBank UK project investigates how

Conference'04, Month 1–2, 2004, City, State, Country.

such repositories might provide opportunities to curate scientific datasets more effectively.

A growing number of institutions are establishing institution-based repositories to manage institutional assets, e.g. e-prints of journal articles and other research publications [16, 42]. The Open Archives Initiative (OAI) [47] architecture provides a suitable basis for interoperability between such repositories and the 'service providers' that provide enhanced access to them. Using the OAI Protocol for Metadata Harvesting (OAI-PMH) [41], institutional repositories are able to expose data to third party service providers, typically aggregators who harvest metadata from multiple repositories, add value and provide user-facing services. The eBank UK project focuses on how institutional repositories might support the curation and dissemination of research data in a similar manner. Experimental data would be made accessible at an early date so that the data can be discovered and made available to both machine and human readers. The aim is to embed this 'deposit and disseminate' process into the workflow of the scientist in as automated way as possible, so as not to add to the burden of their work.

A large number of existing repositories already accept the deposit of certain categories of research data. These include national archives covering particular subject disciplines, e.g. the UK Arts and Humanities Data Service [5], the UK Data Archive [59] and, on a slightly different level, the Atlas Datastore [13] hosted by the Council for the Central Laboratory of the Research Councils (CCLRC). In addition, there are many discipline-specific institutions that curate datasets on behalf of the international scientific community. These include well-known bioinformatics databases like GenBank [28] and the EMBL Nucleotide Sequence Database [22], run respectively by the US National Center for Biotechnology Information and the European Bioinformatics Institute. Others include the PDB Protein Data Bank maintained by the Research Collaboratory for Structural Bioinformatics [52] and the services provided by the Cambridge Crystallographic Data Centre (CCDC) [4, 12]. There has been limited work performed on the use of open archive protocols to publish research data. However, the Reciprocal Net project [57] is a scheme where members of a select distributed network share crystallography results data to form an open and extensible digital collection of molecular structures, either for use by the consortium or for public dissemination and educational use. Also noteworthy is the Crystallography Open Database (COD) [17] where members of the public are invited to upload their crystallography data for open use. The Reciprocal Net and COD projects clearly identify the need for a coherently designed and institutionally based mechanism that would allow any researcher to contribute to the global pool of crystallography data. The eBank UK solution does not seek to replicate discipline-specific initiatives, such as those of CCDC, but to complement them by providing both an institutionally based stage in the cycle of deposit and by allowing for aggregator services to create a variety of added value services.

There are benefits both for the researcher and the institution in depositing their research outputs in a local institution. From the researcher's perspective, the local institution can offer support for the deposit process and can offer additional services such as the management of individual's research output, up-to-date curricula vitae, a single deposit process for multiple reporting requirements. From the institution's viewpoint, a structured repository fulfills legal and funding requirements to store research data, provides a showcase for research outputs, contributes to wider access to research outputs and potentially enhances citation impact [31].

Institutional repositories that expose their metadata for harvesting using the OAI-PMH provide baseline interoperability for metadata exchange and access to data, thus supporting the development of service providers that can add value. Although the provision of added value by service providers is not currently well developed, as e-print archives become deployed more widely within institutions, a number of experimental services are being explored. For example OCLC Research have developed tools that might enable the enhancement of metadata through automatic subject classification and checking the authoritative form of personal and organization names [18].

During the first phase of eBank UK (2003-4) the project investigated the use of digital library technologies for managing datasets, and explored services that might be offered based on the data i.e. linking datasets and journal articles [32, 44]. The project is working in the chemistry domain with the EPSRC (Engineering and Physical Sciences Research Council) funded e-science test-bed Combechem [15] 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 project is working in particular with the crystallography sub-discipline, in the form of the EPSRC National Crystallography Service (NCS) [25], based in the School of Chemistry, University of Southampton. At the end of the first year the project has:

- Gathered requirements from crystallographers both as depositors of research data and as users of research data.
- Developed a demonstrator institutional repository at the University of Southampton [60] for the deposit of crystallography data and metadata that will fulfill requirements of chemists and the local institution, populated with sample metadata relating to research data-sets.
- Developed a demonstrator aggregator service [21] at UKOLN, University of Bath, to harvest metadata about crystallography datasets and scientific papers. Demonstrated how the aggregator might provide an added value service linking research data to papers.
- Developed appropriate schemas to meet the requirements of users of the local repository and aggregator service.
- Demonstrated search interfaces for the local repository and aggregator service.
- Demonstrated a search interface as an embedded service within the PSIgate portal [56] at the University of Manchester.

The aim has been where possible to use existing standards and protocols such as the OAI-PMH, Dublin Core (DC) [19], METS [45]; and to re-use existing open source software as appropriate. The demonstrator is intended to be an exemplar, which will inform discussion of the feasibility of more generic solutions. The rest of this paper will describe these aspects of the project in more detail.

2. REQUIREMENTS

For the development of a demonstrator system, the eBank UK project decided to focus on the sub-discipline of crystallography, as this has a well-defined data creation workflow and a tradition of sharing results data in an internationally accepted standard, the

Crystallographic Information File (CIF) adopted by the International Union of Crystallography (IUCr) [11, 30, 38]. In addition, secondary services like the Cambridge Structural Database (CSD) provide facilities for the acquisition, storage, validation, retrieval, analysis and visualization of small-molecule crystal structures, again mostly available in CIF format [2, 4, 12]. Many crystallographic journals encourage (or require) the submission of structures in CIF format and the CSD acts as an official data depository on behalf of a number of these.

These limited amounts of results data however offer only a partial solution, as the final results dataset is in most cases a small fraction of the data generated during the whole course of the experimental workflow. Also, as we described earlier, publication protocols are time-consuming and existing data centers only deal with a relatively small proportion of the number of structures that have been decoded. The increasing use of high-throughput technologies mean that there is a need for new ways of making such data available [3, 34]. The crystal structure archive created by the eBank UK project was intended to provide access to *all* the different types of data generated during the experimental workflow. It is currently impossible for chemists to obtain datasets produced at earlier stages of the data creation process without the direct co-operation of the research teams that produce them. The NCS is a good case study because of its high sample throughput, state of the art instrumentation, expert personnel and profile in the academic chemistry community.

The main focus for the project was thus to explore the improvement of access to experimental data. This was to be achieved not only by advocating a 'publication at source' philosophy based on open access principles, but also by enhancing the *discovery* process. Specifically, the project was interested in developing service provider models, based on the OAI-PMH architecture adopted by the project. This would primarily be concerned with the development of alternative routes of access and discovery for research data, but the project was also interested in building potential links with more traditional digital library objects, e.g. publications.

Next, we will explore by means of scenarios, some of the requirements for an example aggregator service that links datasets and publications. It is worth emphasizing that the eBank UK project was not seeking to reinvent the subject-specific functionalities developed by IUCr journals or the CSD, but to investigate whether the OAI-PMH could be used to support the creation of institutional repositories designed for crystallographic datasets and the third-party aggregator services that could facilitate the retrieval of datasets published in this way. The scenarios detailed here do not relate to the detailed requirements for the deposit interface and the creation of the crystal structure repository. Instead, a description of the crystallography process, its datasets and the eCrystals archive are provided in Section 3.

The nature of the experimental data and process, combined with aggregator requirements, shape the metadata schema requirements which are discussed in Section 4.

2.1 User Scenarios

2.1.1 Linking from paper to dataset

A crystallography researcher is reading a paper by S. Besli, S. C. Coles, *et al*. that was published in *Acta Crystallographica Section B* in 2002 [8]. This is linked to a report on the structure "2,2-Diphenyl-4,6-cis-oxy(tetraethyleneoxy)-4,6-bis(2,2,2-trifluoroethoxy)cyclotriphosphazatriene". Because the paper was published in an IUCr journal, the researcher knows that she would be able to obtain the final CIF version of the structure from IUCr's Structure Reports Online database or from the CSD. However, because she has already read the supplementary information on the preparation of the compounds, the researcher is interested in acquiring earlier forms of the datasets for reanalysis. From talking with her colleagues, she knows that there is a service called eBank UK that would enable her to tell whether this particular structure was available and which versions of the dataset she could download. She, therefore, points her Web browser at the eBank UK service and uses the simple search interface to look for the first author's name, "Besli". The results page separately lists the paper that she has already consulted and two structure reports; one of them linked to the relevant paper as a "related dataset." Following the link to the dataset related to the paper takes her to the Crystal Structure Report Archive run by the University of Southampton, which gives further information on the structure, some administrative information, and a three dimensional image. Linked to this page are the CIF and a list of data files produced at earlier stages of the experimental workflow. While journals and the CSD currently only make the final results CIF dataset available, eBank UK is able to provide the researcher with access to ALL of the datasets generated during the course of the experimental workflow. The researcher decides which of these earlier datasets she needs to consult and downloads the relevant files for reanalysis. Because she has access rights to all publicly available data linked to the eBank UK system, and because (in this case) the research team that produced the original data have given permission for its free distribution through eBank UK, the researcher can readily download the relevant files for reanalysis. It is perhaps worth noting that in other cases there may be terms and conditions that determine exactly what the researcher is able to do with the data, and information on this would need to be available at the time of download. The issue of access rights is an area that needs more work. Rights metadata have been developed for e-print repositories [27], but it remains to be seen if these will be useful for repositories of data.

Potentially, the eBank UK aggregator could also provide links to those papers that have reused (or cited) previously deposited datasets. If our hypothetical researcher used the downloaded data to produce a new structure report and paper, these could be deposited in her own institution's repository. If both the paper and the structure report provide a citation to the URI used by the University of Southampton's Crystal Structure Report Archive, it might be possible for the eBank UK aggregator to match the links to provide a link from the new structure report and paper to the old structure. In this case, on searching again for the author "Besli", the metadata for the structure report discussed above would be provided with "related paper" link to the original article by Besli, *et al*., and the new paper and structure reports produced by the researcher and her team.

2.1.2 Searching for datasets

A PhD student is looking for some crystal structures produced by a research group based at another UK university. He knows the general type of compound that he is looking for, but does not know the exact formula or the IUPAC name. He elects to search first for these structures through the eBank UK aggregator and connects to the service. He searches for the name of the research

group head in the "author" field, limiting his search by the compound class "Organic" and by date to retrieve only the most recent structures submitted to the system. The results provide two lists, one for crystal structure reports, the other for publications. Because the student has specified the type of compound class, the search only retrieves details of 15 structure reports. He browses the results, checking the formula and IUPAC name fields for information on the chemical makeup of the crystals. He is interested in just five of the structures, so for each of these, he links to the locally hosted structure report repository to view the more detailed metadata for each one and to see exactly what data is available for download. If he wishes to download any of these datasets for reanalysis, he may have to have to fulfill the authentication requirements of the local system (data provider).

To reiterate, for the pilot aggregator the eBank UK project was not trying to emulate the more sophisticated search functionality of the CSD or even the IUCr Structure Reports Online database. The latter permits searching on bibliographic-type metadata and the full-text of papers. CSD provides more advanced ways of searching for chemical information. CCDC's free CIF depository request service enables retrieval by CCDC deposit codes or the bibliographic citation of a related publication. It would be possible for users to search first the CSD for specific chemical features and then use the information retrieved (e.g., CCDC codes or bibliographic information) to search the eBank UK service. The OAI publication model, when applied to research data, will promote the growth and potential added value of subject-based services and therefore enhance the service provided by CCDC.

3. DEMONSTRATORS

3.1 The Crystallography Workflow

A crystallography experiment consists of a series of processes that ultimately result in the determination of a crystal structure, expressed in the form of the CIF file (referred to above). A number of well-defined stages, either measurement or analytical, are carried out in sequence. At each stage an instrument or computational process produces an output, commonly saved as one or more data files, where the output of one stage constitutes the input to the next stage. A number of stages are readily identifiable which, independent of procedure adopted, software used and specific data being examined, always results in files of the same content and format.

This results in the crystallography workflow, or pipeline, outlined in Figure 1, which depicts the individual processes and maps them onto the files generated (this is not intended to be a convention and merely describes the NCS workflow). For publication purposes the CIF file (readable using the ENCIFER software [23]) is currently perceived to be the final result of a crystallographic experiment, but in eBank UK value is added in order to enable linking and aggregation through OAI-PMH. For example, CHECKCIF [37] is a web-based structure validation program, which produces an output file formatted as HTML and represents the *validation* stage of the process. In addition a CML (Chemical Markup Language) [46] file is generated that enables the exchange of this chemical structure information to be automatic in addition to being platform and software independent. The INChI identifier [39] is also generated to be a unique text representation of the molecule and therefore assist in linking and aggregating processes.

In each experiment, the process relates to the determination of one structure, that is the determination of both the molecular connectivity and the packing arrangements between molecules in the crystal examined. The output files have varying formats representing information about the molecule, from images to highly-structured data expressed in textual form, and the file extension names are explicit in the field.

At one level, the files themselves can be considered to contain metadata about the molecules or the experiment itself, e.g. validation parameters which express the confidence in the accuracy of the result. This can have interesting consequences for the modeling and requirements discussions, as outlined later.

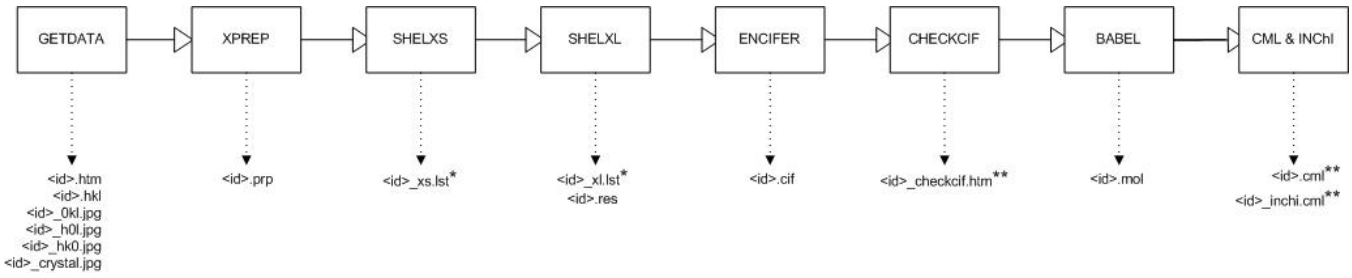

Figure 1 A typical Crystal Structure Determination Workflow

3.2 The Local Repository interface

A test data provider repository [60] was created at the University of Southampton and populated with test metadata. The repository supports the deposition of data from stages of the data creation process and workflow, the creation of metadata either by the depositor or automatically, and the browsing and searching through its own local web interface. Users can search the repository using a range of bibliographic and chemical parameters, and browse by date, creator name or class of compound.

The schema employed in the archive has the unique file extensions associated with the appropriate part of the process and therefore is able to 'recognize' a particular file when it is presented. This enables a simple deposition process whereby the depositor supplies core bibliographic information and some chemical metadata, which is marked up as a combination of regular and qualified Dublin Core, along with a ZIP file containing the digital output from the experiment. In addition to making the files available, 'quality indicators' are extracted from a number of these files. This key information is presented alongside the files for download, the author input metadata and a rendered version of the CML file, which is made interactive through the use of an applet. When a stylesheet is applied to this data an entry in the archive is displayed as shown in Figure 2. The crystal structure repository uses a specially enhanced version of the EPrints.org software developed at the University of Southampton [24].

3.3 The Aggregator

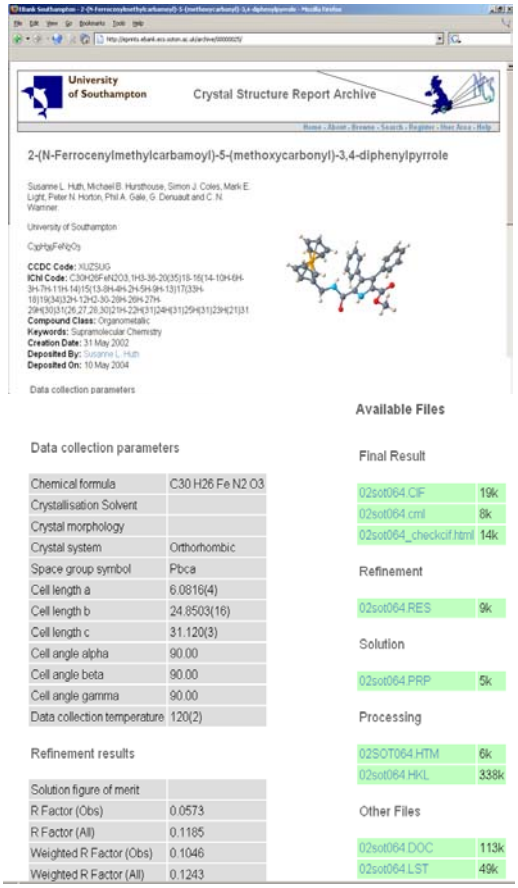

Figure 2 The Repository Web Interface

This demonstrator [21] harvests metadata from the crystal repository and implements the searching functionality described in the scenarios, cross-searching metadata on research data and publications. The search was also embedded into an external web site at the PSIgate service.

4. METADATA, MODELS AND SCHEMAS

The OAI-PMH model of metadata dissemination provides for the distribution of metadata records in different formats. The metadata is intended to be used by service providers in the OAI-PMH model as a basis for building value-added services [47]. Models of service providers built on harvested metadata are still evolving, and a 2003 survey suggests that there is still limited experience with the development of service providers [10]. Within the general aim of opening up access to research data by improving dissemination routes for the metadata, the eBank UK aggregator demonstrator focused on one potential added value service. This service is located within the vision for the scholarly knowledge cycle [43], and seeks to provide an integrated search linking publications with related experimental data. The user scenarios outlined earlier typify the hypothetical use of such a service by a researcher in the field of crystallography.

The design of the schemas that define the metadata formats that could be distributed by the local crystallography OAI-PMH repository was thus partly driven by consideration of the requirements for our prototype service provider. As seen in Figure 2, the local repository (i.e. the data provider) presents as a web page a jump-off page with several fields of information extracted from the files. User requirements indicated that for the purposes of the aggregator demonstrator, a much smaller subset of the metadata would be sufficient to provide a basis for cross-searching and to flag the presence of relevant resources to the end-user. The interested user could then peruse further experimental details and access data files by following web links to the local repository.

In other words, service provider searches would simply be indicative of the experimental data available; the gateway to the data itself would be located at the local crystallography repository. With one important exception, all the information that was required for exposure in the metadata related generally to the whole of the experimental process. Thus the (human) creators of all the data files relating to one experiment deposited within the repository were common across all data files. Furthermore, it was expected that these creators would correspond to the authors of the relevant publications in the cross search at the service provider end. The subject (or topic) of the experiment (and therefore of all the files) was the one crystal structure (molecule) that was being analyzed during a specific run of the crystallography process. It was anticipated by the users that although depositors would be allowed to add files incrementally, the experiment would only be exposed to the outside once the whole experiment was available, therefore there would only be one common date available. The description of the data files, in the main, could be generalized to model and describe them as one single collective resource.

4.1 Using Dublin Core

To design the metadata schema, it was decided to use DC without qualification as a basis and expand to an extended version as required. The reasons for this firstly that the unexpanded version

would be necessary to fulfill the OAI-PMH minimal requirements. Secondly, those elements from DC that could be retained, together with the qualified elements (mapped to the unqualified form), would provide a convenient basis for cross searching within service providers aggregating other DC metadata.

Two areas in particular that required the use of the extended form of DC can be highlighted. First, the subject of the experiment is the crystal structure of a molecule, which is described by various methods. These include official names, chemical formulae and a more recently formulated identifier, the INChI [39], as well as codes that identify the crystal structure entry in the databases mentioned previously. These different names for the molecule can be considered to be taken from different vocabularies.

The Dublin Core Metadata Initiative (DCMI) recommends the following for including terms from vocabularies in encoded XML, suggesting: "*Encoding schemes* should be implemented using the 'xsi:type' attribute of the XML element for *property*". As there are not as yet any designated names for chemistry vocabularies, the project defined some eBank terms as designators of the type of vocabulary being used to describe the molecule. Thus an empirical formula would be expressed in the metadata record as:

<dc:subject xsi:type="ebankterms:EmpiricalFormula">C27H48</dc:subject>

In the longer term, it would be desirable if standardized methods were agreed within the chemistry community for defining the terms that designate a specific naming convention, using namespaces to support XML processing. The eBank terms should be considered to be no more than placeholders until official ones become available. Work in this area has already commenced within internationally recognized bodies in chemistry [62]

Secondly, the users had prescribed that as part of the indicative information provided to the service provider, it would be desirable to state the type of files available. Rather than individual file types, the chemistry users thought that stage information would enable a researcher using an aggregator system to form an opinion regarding datasets of interest and range of data available for a particular experiment. Files were designated as belonging to the initialization, collection, processing, solution, refinements or CIFDataset stage. Once again, a schema was used to define these terms as the EbankDatasetType vocabulary.

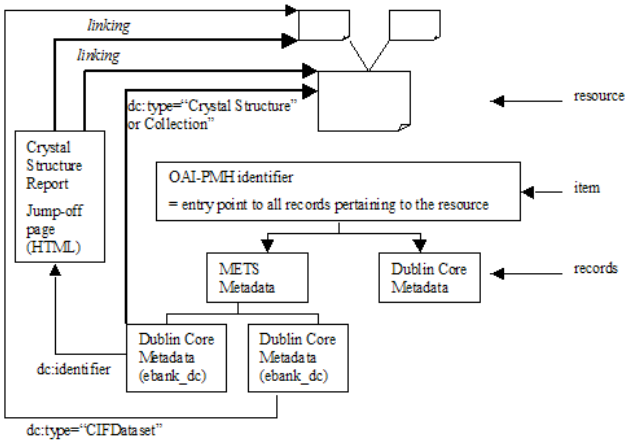

Figure 3 Representation and linking of resources

However, stage information could not be generalized across the data files as had been done hitherto, since the various stages related to different files. This introduced a further layer of complexity to the exposed metadata. The proposed Dublin Core abstract model [54] stipulates a one-to-one relationship between a description and a resource. The stage for the file is a property of the individual file, which is a resource that can in itself be described separately from the *collection* of data files belonging to the experiment. The proposed abstract model for DC introduces the notion of a *description set* for descriptions of related resources. The guidelines for expressing DC in XML (at the time of designing the eBank UK schema) however stipulated that the instantiation of a DC description within a record should relate to a description of only *one* resource. It is likely that the guidelines for encoding DC in XML will change [55]. However at the time it was necessary to produce multiple DC records, one for the general description relating to the collection of data files, and minimal ones describing the stages. The latter consisted of simply an identifier for the stage and the stage type chosen from the EbankDatasetType defined in the schemas. Figure 3 (adapted from [61]) shows the relationship between the metadata records and the resources within the context of the OAI-PMH model.

To exchange the set of DC records within the OAI-PMH, a wrapper schema is required since an XML record must have an outer element. The choices available to the project were to define its own wrapper to enclose the Dublin Core XML records, or to consider the use of existing packaging standards intended to disseminate related sets of metadata.

The project chose to look at METS [45]. Although the choice was not made following a detailed analysis of all alternative packaging formats (which was beyond the scope of the initial phase of the project and the time available) it was not completely arbitrary. METS was freely available, it has inherent support for the inclusion of DC metadata, and it also carries currency in the digital-library community.

Whilst the application of complex object modeling and description in the application area is at very early stages, this is an area that deserves greater attention and study. Clearly, there may be scope for other data-file or stage-specific information (e.g. provenance metadata which describes the origin of the data) that would be attached to a specific stage, rather than to the whole experimental process. Descriptions of access restrictions (or other requirements, for example, visualization tools for viewing the data) may similarly need to be more granularly related at the stage or file level. Thus although the use of packaging has been simplistic, it is evident that the treatment of research data as a complex digital object would benefit from consideration of approaches being advocated in the digital library community.

It should be stated that any recommendation requires more than the superficial use of packaging standards that the project has made so far, and should involve an evaluation of the alternatives against wider requirements. Requirements in the application domain are as yet largely unknown or at best unspecified.

Another aspect of granularity relates to that of identification. As explained, the collection of data files was treated for the main part as one resource. There are as yet no standardized means of identification of data within the crystallography community, other than that provided by the large databases, which for example provide codes to access specific molecular entities. Rather than

pre-empt any decisions in this regard, identification and access to the experimental data was achieved using the HTML entry points provided in the local repository. Other scientific communities are at a more advanced stage of agreeing systems for identification and resolution. Once again, a better understanding of the use of crystallography data and its dissemination is required. Scenarios of use could be built surrounding the re-use of data sets in subsequent experiments. This would have implications for the description of the later experiments and the methods used to reference the data sets available in OAI repositories. However such scenarios are still to be discussed with the chemistry data producers and the wider chemistry community.

4.2 Data Modeling

Reconciling the model of the experimental data process with the OAI-PMH model of metadata dissemination, and that of Dublin Core was not always straightforward. This was due in part to communication difficulties, rather than inherent problems in the various models. The use of terms such as data, metadata and record within an interdisciplinary team must be made judiciously and after negotiation and explanation of the context within which they are used, otherwise misunderstandings can easily arise. For example, the duality of the role of data contained within the data files, which at times was part of the data, but at other times fulfilled the role of exchanged metadata, or simply additional information displayed to a user, made discussion and reconciliation of views difficult, particularly at the initial stages when a common ground of understanding between the different partners was still being established.

Furthermore, the so-called 'jump-off page' or 'splash-page', commonly presented in data provider repositories as browsable HTML pages, is not recognized as a prominent entity in the OAI-PMH model. In practice, the jump-off page is often used as an entry point to resources present in a repository. Often, browsing and access facilities are an integral part of both discovery and access to OAI-PMH repositories. In some instances, locating the resource requires passage through the jump-off page, since its location is given as a proxy identification for the resource itself within the metadata. In our particular application domain, the users considered the jump-off page as a pivotal point of access to the data, containing links to the individual data files. Moreover, discussion with the chemistry users revealed that they considered this page to be a form of 'report', a distillation of the experiment, with detailed information of importance to the specialized end-user. This page contained in essence some of the metadata that was required and essential from a machine-processing point of view. However, from a human end-user viewpoint the browsable HTML pages are clearly of professional importance, showcasing as they do the quality and extent of the work presented through the local repository. In deciding system requirements, it is useful to make a clean separation between the machine-processable components and those intended for human consumption. It is important to ensure that discussion maintains a clear distinction between metadata-resource relationship models and the human-oriented presentation as, for example, jump-off pages or reports.

4.3 Caveats

It should be emphasized that this is a work in progress and the above discussion is based on the experience gained and progress made up to the end of the first phase of the project as of September 2004. At this stage the schemas are by no means intended to be normative; they have provided a means to explore some service provider issues and illustrate some of the decisions that need to be taken by the community to provide specialized crystallography OAI services. These include decisions regarding the use of identifiers and the granularity at which they are applied, and the definition of terminology for describing molecules and experimental characteristics. It would be foolish to make any claims to having explored all the metadata requirements This notwithstanding, the demonstrator and fledgling schema have been sufficient to fuel interest and motivate collaboration.

5. RELATED WORK

5.1 CMCS

The Collaboratory for the Multi-Scale Chemical Sciences (CMCS) [14] is a metadata-aware system that supports information exchange between chemistry sub-disciplines (although not in crystallography). Its functions are wider than simple metadata exchange, supporting: metadata querying, metadata generation, tools for data and metadata management, metadata mapping, email alerts and data visualization. CMCS has developed standards for data and metadata description.

Dublin Core was used as a foundation for description of the data: metadata definitions in CMCS use DC basic elements, and element refinements. Particular use was made of the following elements to capture the relationship of scientific data sets to other data and for recording traceability: *Is Version Of*, *Has Version*, Is *Replaced By* and *Has References*. CMCS offers a core schema with metadata elements to record chemistry-specific data (e.g. species name and formula) and to record relationships of data to projects was defined.

Despite the slightly different aims and application areas of the CMCS, common concerns can be identified. Mapping of metadata was used in order to use DC within the system whilst allowing researchers to use more familiar terminology or chemical science schemas within their data. The project recognized that "Enforcing metadata standards across multiple chemistry communities would not be pragmatic and would alienate scientists. Instead we are providing guidance to users capturing metadata to simplify mapping within CMCS" [50].

5.2 Complex objects and the OAI-PMH

There is growing interest in, and encouragement of, the use of complex object descriptions and their dissemination using the OAI-PMH. A number of metadata standards that accommodate the description of complex digital objects are available.

METS is an initiative of the library community, maintained by the Library of Congress Network Development and MARC Standards Office. METS recognizes that describing digital objects requires an increasingly complex series of metadata descriptions - administrative, structural and technical metadata, for example. A review of METS and examples of its applications featured in a recent issue of the journal *Library Hi Tech* [55]

Bekaert, *et al*. highlighted the relevance of MPEG-DIDL [40] to the digital library community [6]. MPEG-21 is an ISO-approved standard, and its framework provides a well-defined data model for complex digital objects, as well as an XML Schema for representing compliant digital objects. The standard has been used to represent and disseminate complex digital objects from the collection at the Los Alamos National Laboratory Research

Library [7]. More recently, complex object modeling and description has been proposed as a solution to some as yet unsolved issues in the OAI-PMH infrastructure, namely the need to transfer digital content from one data repository to another for preservation purposes, and the requirement to reliably access content itself (not simply the metadata) [61].

The eLearning community has also addressed the challenge of describing complex digital resources by defining metadata standards. Two leading examples are IMS Content Packaging [35] and the Shareable Content Object Model (SCORM) [1].

These various approaches and standards for the description of complex digital resources may offer different facilities and features. Their applicability to the description of complex scientific data is still to be evaluated. Early indications of similarities in requirements between different data domains must be backed up by more detailed analysis and experimentation, particularly in the contexts of using exchanged metadata to aid the access and discovery of resources.

5.3 Scientific datasets and identification

As mentioned, there is as yet no standard agreed method for identifying and locating crystallography data outside of the mainstream databases. Other scientific communities are at a more advanced stage of establishing such mechanisms.

The use of DOIs in the scientific field has been recently presented by Paskin [51], and in particular two projects were highlighted as providing interesting DOI applications with science data. One project in Germany [9] has developed a pilot that applies DOIs to climate data and uses these identifiers for citation and re-use in the long-term referencing of primary data. The above review highlighted the outstanding issue of granularity in the assignment of identifiers, relating it to the need to consider the functional requirements in their assignment: "DOIs could logically be assigned to every single data point in a set; however in practice, the allocation of a DOI is more likely to be to a meaningful set of data following the indecs Principle of Functional Granularity: identifiers should be assigned at the level of granularity appropriate for a functional use which is envisaged" [51].

The Life Science Identifier (LSID) project has recently emerged in the biosciences and is maintained by the Interoperable Informatics Infrastructure Consortium (I3C) [36].

6. FUTURE WORK

During its first phase, UK focused exclusively on chemistry and in particular on crystallography. During its second year of funding, the project will seek to collaborate more closely with other scientific domains whilst continuing to work with the wider crystallography community to validate initial results.

The project will explore the feasibility of applying the eBank UK architecture, data models and schema data in other sub-disciplines of chemistry and the physical sciences. It will support consensus building within the digital library and scientific communities on the development of a generic data model and metadata schema for scientific data. One aspect of this work will be to consider the feasibility of developing repository software that can easily be configured for a variety of scientific data.

Discussion with publishers indicates that the eBank UK approach is a promising solution to the current publication bottleneck problem. The project will build on its initial collaboration with the two principal international crystallographic organizations (and also publishers), the International Union of Crystallography (IUCr) and the Cambridge Crystallographic Data Centre (CCDC). The aim is to integrate the eBank UK approach into crystallography related publications so that in the future it will become an accepted form for publishing crystal structures.

In order to progress the added value service demonstrated by eBank UK, that is linking data and publications, much more work needs to be done to reach agreement on a common approach to citation of data within publications. The project will address this issue by exploring the use of persistent identifiers for e-research data including "generic" and intra-domain identifier systems e.g. the InCHI. Related to this work will be investigation of the use of resolution tools, i.e. OpenURL, to facilitate linking from primary data to peer-reviewed published articles and to explore the potential for additional context sensitive linking.

The project will evaluate the pedagogical benefits of enhancing access to primary e-research data through the eBank UK service. This will be done by providing access to research data within e-learning materials in the context of a taught postgraduate course in chemistry. To facilitate this study, access to primary research data outputs will be embedded in learning materials in a number of ways e.g. through links in reading lists, through analytical problems, through embedded links in portals such as that demonstrated by the project in PSIgate.

Populating institutional repositories has proved to be a significant issue in the context of ePrints. EBank UK will explore further possibilities for automated creation of metadata for datasets. Within the Combechem project work has been progressing on both data and metadata acquisition in a smart lab context (c.f. the Smart Tea project [58]). In phase two, the project will investigate how to exploit data and metadata acquired or derived by such systems.

7. CONCLUSION

The potential benefits of institutional repositories to the wider community beyond the institution rely on external service providers exploiting the content of multiple repositories, both institutional and subject based. Effective services will be built by coherent aggregation of content from distributed networks of repositories. Within the scientific domain there are already a number of existing activities that might be integrated into such an open access architecture. It seems likely there will be no 'single solutions' rather a network of interoperable solutions

Within the eBank UK project we have explored the implementation of a single repository and associated aggregator service whilst acknowledging by our collaborations that future services, even within a particular discipline, will rely on a well-structured workflow connecting a network of multiple repositories interfacing with service providers and possibly other components of the information environment.

Our development activities and discussions during the first phase of the project have highlighted the challenges and complexities of working in a cross-domain area, both in terms of understanding the nature of the chemistry data and its management, but also in terms of the "landscape views" of different communities. Involvement of subject specialists is vital to ensure the data model fits with specialist scientific data. In converse, in order to build interoperable solutions it is essential to raise awareness of digital

library technologies in specialist areas of informatics, a point that has been made already in relation to computer scientists' contribution to eScience [29]. There is a need for professionals from different domains to work effectively together, seeking solutions for specialist areas whilst maintaining awareness of the benefits of wider interoperability.

8. ACKNOWLEDGEMENTS

The eBank UK project is funded by JISC under the Semantic Grid and Autonomic Computing Programme. The authors are indebted to the other project members who have contributed to the work described in this paper: J.G. Frey, M. Hursthouse (School of Chemistry, University of Southampton), C. Gutteridge, (School of Electronics and Computer Science, University of Southampton), L. Lyon, A. Powell (UKOLN, University of Bath), J. Blunden-Ellis, P. Meehan (PSIgate, University of Manchester).

9. REFERENCES

[1] Advanced Distributed Learning. The Sharable Content Object Reference Model (SCORM). Retrieved January 27, 2005, from: http://www.adlnet.org/

[2] Allen, F. H. The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Cryst*., B58, (2002), 380-388.

[3] Allen, F. H. High-throughput crystallography: the challenge of publishing, storing and using the results. *Crystallography Reviews*, 10 (2004), 3-15.

[4] Allen, F. H., Bellard, S., Brice, M. D., Cartwright, B. A., Doubleday, A., Higgs, H., Hummelink, T., Hummelink-Peters, B. G., Kennard, O., Motherwell, W. D. S., Rodgers, J. R. and Watson, D. G. The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Cryst*., B35, (1979), 2331-2339.

[5] Arts and Humanities Data Service. Retrieved January 27, 2005, from: http://www.ahds.ac.uk/

[6] Bekaert, J., Hochstenbach, P. and Van de Sompel, H. Using MPEG-21 DIDL to represent complex digital objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine*, 9, 11 (November 2003). doi:10.1045/november2003-bekaert

[7] Bekaert, J., Balakireva, L., Hochstenbach, P. and Van de Sompel, H. Using MPEG-21 DIP and NISO OpenURL for the dynamic dissemination of complex digital objects in the Los Alamos National Laboratory Digital Library. *D-Lib Magazine*, 10, 2 (February 2004). doi:10.1045/february2004-bekaert

[8] Besli, S., Coles, S. J., Davies, D. B., Hursthouse, M. B., Kiliç, A., Mayer, T. A. and Shaw, R. A. Structural investigations of phosphorus-nitrogen compounds. 5. Relationships between molecular parameters of 2,2-diphenyl-4,6-*cis*-oxytetra(ethyleneoxy)-4,6-R_2-cyclotriphosphazatrienes (R = Cl, OCH_2CF_3, OPh, OMe, NHPh, $NHBu^t$) and substituent basicity constants. *Acta Cryst*., B58, (2002), 1067-1073.

[9] Brase, J. Using digital library techniques: registration of scientific primary data. In *Research and advanced technology fir digital libraries: 8th European Conference, ECDL 2004*. Springer, Berlin, 488-494.

[10] Brogan, M. A. *A survey of digital library aggregation services*. Digital Library Federation, Council on Library and Information Resources, Washington D.C., 2003. Retrieved January 27, from: http://www.diglib.org/pubs/brogan/

[11] Brown, I. D. and McMahon, B. CIF: the computer language of crystallography. *Acta Cryst*., B58, (2002), 317-324.

[12] Cambridge Crystallographic Data Centre. Retrieved January 27, 2005, from: http://www.ccdc.cam.ac.uk/

[13] CCLRC Atlas Datastore. Retrieved January 27, 2005, from: http://www.e-science.clrc.ac.uk/web/services/datastore

[14] Collaboratory for Multi-Scale Chemical Science (CMCS). Retrieved January 27, 2005, from: http://cmcs.ca.sandia.gov/

[15] Combechem project. Retrieved January 27, 2005, from: http://www.combechem.org/

[16] Crow, R. *The case for institutional repositories: a SPARC position paper*. Scholarly Publishing & Academic Resources Coalition, Washington, D.C., 2002. Retrieved January 27, 2005 from: http://www.arl.org/sparc/IR/ir.html

[17] Crystallography Open Database. Retrieved January 27, 2005, from: http://www.crystallography.net/

[18] Dempsey, L., Childress, E. R., Godby, C. J., Hickey, T. B., Vizine-Goetz, D. and Young, J. Metadata switch: thinking about some metadata management and knowledge organization issues in the changing research and learning landscape. In Shapiro, D., ed., *LITA guide to e-scholarship* [working title]. American Library Association, Chicago, IL (forthcoming). Preprint retrieved January 27, 2005, from: http://www.oclc.org/research/publications/archive/2004/dempsey-mslitaguide.pdf

[19] Dublin Core Metadata Initiative. Retrieved January 27, 2005, from: http://dublincore.org/

[20] eBank UK project. Retrieved January 27, 2005, from: http://www.ukoln.ac.uk/projects/ebank-uk/

[21] eBank UK aggregator demonstrator. Retrieved January 27, 2005, from: http://eprints-uk.rdn.ac.uk/ebank-demo/

[22] EMBL Nucleotide Sequence Database. Retrieved January 27, 2005, from: http://www.ebi.ac.uk/embl/

[23] enCIFer software. Retrieved January 27, 2005, from: http://www.ccdc.cam.ac.uk/free_services/encifer/

[24] EPrints.org. Retrieved January 27, 2005, from: http://www.eprints.org/software.php

[25] EPSRC National Crystallography Service. Retrieved January 27, 2005, from: http://www.soton.ac.uk/~xservice/

[26] Frey, J. G., Bradley, M., Essex, J. W., Hursthouse, M. B., Lewis, S. M., Luck, M. M., Moreau, L. A. V. M., De Roure, D. C., Surridge, M. and Welsh, A. H. Combinatorial chemistry and the Grid. In Berman, F., Fox, G. and Hey, A. J. G., eds., *Grid computing: making the global infrastructure a reality*. Wiley, Chichester, 2003, 945-962.

[27] Gadd, E., Oppenheim, C. and Probets, S. RoMEO studies 6: rights metadata for open archiving. *Program*, 38, (2004), 5-14.

[28] GenBank. Retrieved January 27, 2005, from: http://www.ncbi.nlm.nih.gov/Genbank/

[29] Gray, J. and Szalay, A. *Where the rubber meets the sky: bridging the gap between databases and science*. Technical Report MST-TR-2004-110, Microsoft Research, Redmond, WA, December 2004. Retrieved January 27, 2005, from: http://research.microsoft.com/research/pubs/view.aspx?tr_id=815

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

[31] Harnad, S., Brody, T., Vallières, Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H. and Hilf, E. R. The access/imact problem and the green and gold roads to open access. *Serials Review*, 30, (2004), 310-314. Retrieved January 27, 2005, from: http://eprints.ecs.soton.ac.uk/9939/

[32] Heery, R., Duke, M., Day, M., Lyon, L., Hursthouse, M. B., Frey, J. G., Coles, S. J., Gutteridge, C. and Carr, L. A. Integrating research data into the publication workflow: the eBank UK experience. PV-2004: Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data (Frascati, Italy, October 5-7 2004). Retrieved January 27, 2005, from: http://www.ukoln.ac.uk/projects/ebank-uk/dissemination/

[33] Hey, T. and Trethethen, A. The data deluge: an e-science perspective. In Berman, F., Fox, G. and Hey, A. J. G., eds., *Grid computing: making the global infrastructure a reality*. Wiley, Chichester, 2003, 809-824.

[34] Hursthouse, M. B. High-throughput chemical crystallography (HTCC): meeting and greeting the combichem challenge. *Crystallography Reviews*, 10, (2004), 85-96.

[35] IMS Content Packaging Specification. Retrieved January 27, 2005, from: http://www.imsglobal.org/content/packaging/

[36] Interopeable Informatics Infrastructure Consortium (I3C). Retrieved January 27, 2005, from: http://www.i3c.org/

[37] IUCr checkCIF. Retrieved January 27, 2005, from: http://checkcif.iucr.org/

[38] IUCr Crytallographic Data File. Retrieved January 27, 2005, from: http://www.iucr.org/cif/

[39] IUPAC Chemical Identifier (InChI). Retrieved January 27, 2005, from: http://www.iupac.org/projects/2000/2000-025-1-800.html

[40] ISO/IEC 21000-2:2003. Information Technology -- Multimedia framework (MPEG-21) -- Part 2: Digital Item Declaration. International Organization for Standardization, Geneva, 2003.

[41] Lagoze, C., Van de Sompel, H., Nelson, M. and Warner, S., eds., The Open Archives Protocol for Metadata Harvesting, v. 2.0, 14 June 2002. Retrieved January 27, 2005, from: http://www.openarchives.org/

[42] Lynch, C. A. Institutional repositories: essential infrastructure for scholarship in the digital age. *ARL Bimonthly Report*, 226 (2003). Retrieved January 27, 2005, from: http://www.arl.org/newsltr/226/ir.html

[43] Lyon, L. eBank UK: building the links between research data, scholarly communication and learning. *Ariadne*, 36 (July 2003). Retrieved January 27, 2005, from: http://www.ariadne.ac.uk/issue36/lyon/

[44] Lyon, L., Heery, R., Duke, M., Coles, S., Frey, J., Hursthouse, M., Carr, L. and Gutteridge, C. eBank UK: linking research data, scholarly communication and learning. Third UK e-Science Programme All Hands Meeting (AHM 2004) (Nottingham, UK, August 31 - September 3 2004). Retrieved January 27, 2005, from: http://www.allhands.org.uk/proceedings/papers/237.pdf

[45] Metadata Encoding and Transmission Standard (METS). Retrieved January 26, 2005, from: http://www.loc.gov/standards/mets/

[46] Murray-Rust, P., Rzepa, H.S. and Wright, M. Development of Chemical Markup Language (CML) as a system for handling complex chemical content. *New J. Chem.*, 25, (2001), 618-634.

[47] Open Archives Initiative (OAI). Retrieved January 27, 2005, from: http://www.openarchives.org/

[48] Open Archives Initiative. Frequently asked questions. Retrieved January 27, 2005, from: http://www.openarchives.org/documents/FAQ.html

[49] Open Language Archives Community (OLAC). Retrieved January 26, 2005, from: http://www.language-archives.org/

[50] Pancerella, C., Hewson, J., Koegler, W., Leahy, D., Lee, M., Rahn, L., *et al*. Metadata in the Collaboratory for Multi-scale Chemical Science. DC-2003: the 2003 Dublin Core Conference (Seattle, WA, USA, 27 September - 2 October 2003). Retrieved January 27, 2005, from: http://purl.oclc.org/dc2003/03pancerella.pdf

[51] Paskin, N. Digital Object Identifiers for scientific data. 19th International CODATA Conference (Berlin, Germany, 7-10 November 2004). Retrieved January 27, 2005, from: http://www.doi.org/topics/041110CODATAarticleDOI.pdf

[52] PDB Protein Data Bank. Retrieved January 27, 2005, from: http://www.rcsb.org/pdb/

[53] Powell, A. DC Architecture WG meeting report, DC2004. Mail to DCMI Architecture Group mailing list <dc-architecture@jiscmail.ac.uk>, 28 October 2004. Retrieved January 27, 2005, from: http://www.jiscmail.ac.uk/cgi-bin/webadmin?A2=ind0410&L=dc-architecture&T=0&F=&S=&P=4565

[54] Powell, A., Nilsson, M., Naeve, A. and Johnston, P. *DCMI abstract model*. DCMI Working Draft, 8 December 2004. Retrieved January 27, 2005, from: http://www.dublincore.org/documents/abstract-model/

[55] Proffitt, M., Pulling it all together: use of METS in RLG cultural materials service. *Library Hi Tech*, 22, (2004), 65-68. doi:10.1108/07378830410524503

[56] PSIgate. Retrieved January 27, 2005 from: http://www.psigate.ac.uk/

[57] Reciprocal Net. Retrieved January 27, 2005, from: http://www.reciprocalnet.org/

[58] Smart Tea project. Retrieved January 27, 2005, from: http://www.smarttea.org/

[59] UK Data Archive. Retrieved January 27, 2005, from: http://www.data-archive.ac.uk/

[60] University of Southampton, Crystal Structure Report Archive. Retrieved January 27, 2005, from: http://ecrystals.chem.soton.ac.uk/

[61] Van de Sompel, H., Nelson, M. L., Lagoze, C. and Warner, S. Resource harvesting within the OAI-PMH framework. *D-Lib Magazine*, 10, 12 (December 2004). doi:10.1045/december2004-vandesompel

[62] XML Data Dictionaries in Chemistry. Retrieved January 27, 2005, from: http://www.iupac.org./projects/2002/2002-022-1-024.html