Metadata-based Discovery: Experience in Crystallography

Monica Duke

UKOLN, University of Bath

Abstract

Facilitating discovery is an aspect of curation that has been addressed by the eBank UK project. This paper describes the metadata and associated issues considered when working within the chemistry sub-discipline of crystallography. A metadata-mediated discovery model has been implemented to aid the dissemination and sharing of research datasets; although the specific characteristics of crystallography were the main driver in determining the metadata requirements, consideration was also given to cross-disciplinary interaction and exchange. An overview of the metadata profile that has been developed is provided, against the background of the project.

1. Discovery in the curation life-cycle.

Digital curation can be viewed as the active management of data over the life-cycle of scholarly and scientific interest, and is the key to reproducibility and re-use. [1] Within this view, curation is seen as encompassing activities that support the *immediate* use of data as well as its long-term preservation.

Metadata for resource discovery and retrieval is considered to play an important role in this process. Metadata-mediated discovery relies on the description of the data in such a manner as to support discovery services that match a search requirement against some characteristic of the data. Metadata is intended to promote contemporary discovery and use, as well as future unintended uses.

The eBank UK project has addressed a perceived shortfall in the current publication and dissemination process in the field of crystallography, by designing and implementing an open access repository and improving dissemination routes for the associated metadata.

1.1 Sharing and Discovering Crystallographic Data

McMahon [2] provides a historical overview of electronic metadata management in the field of crystallography, from the publisher's perspective. After referring to the processes around bibliographic metadata management in scientific journals, (which are common across publishers working in an electronic environment), the overview deals with the scientific metadata in crystallography, and particularly the standard data exchange format CIF [3], which the community developed in the nineties. The paper highlights the tradition of depositing data in support of published articles in the field of crystallography, and the success achieved by policies to archive data in the agreed uniform format. The CIF format is itself rich in metadata characterizing the data and results derived from it.

A number of repositories of crystallography data are in existence or in development with the aim of allowing human users to query, discover and access their content. They vary in their subscription access model (free, partial or fully fee-based), and in the range of additional services offered e.g. visualization software. (for two example repositories see [4] and [5]). What is common is that all these repositories restrict searching capabilities to the entry of search criteria on a web form as the sole point of discovery, (with an additional email request required to access the data in some cases).

Whilst [2] suggests a number of potential ways of sharing queries across these repositories, and points to example technologies that may fulfil that role in the future, it is clear that an integrated method of discovery across crystallography resources is not yet a reality, despite the use of the common CIF format for storing results data.

1.2 The contribution of the eBank UK project.

The eBank UK approach to data sharing has been two-pronged: the provision of a human-browsable repository of data, and the exposing of descriptive metadata using an internationally agreed protocol, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [6].

The details of the archive and the chosen protocol are described in other contributions submitted to this workshop, or elsewhere [7], [8], [9], [10] and shall not be repeated here. Briefly however, datasets created during a

crystallographic determination are deposited into an archive. During deposition, a number of metadata items are captured, either by automatic extraction from the data or through input by the depositor. The data is then made available through a web interface and can be searched or browsed. Each entry displays the key metadata and gives access to the underlying data. Additionally, the metadata can be accessed through a machine interface using a harvesting model. Third parties are able to issue http requests to retrieve the XML-formatted metadata, which can then be stored locally, integrated and cross-queried with any other metadata available.

The advantage of this approach is that it opens up the metadata to be re-used and incorporated into multiple services, thus potentially generating new pathways to the data. It is not only human users who can query, discover and access the content. The requests for metadata need not be initiated by a human user but can be delegated to an automated process. A machine-processable format of the metadata is provided to promote flexibility in its use and diversity in the discovery services that can be layered above. In particular, one focus of the project was to explore the link between digital library technologies (such as the OAI-PMH) and the sharing of scientific data and metadata, and potential connections with the published literature as a method of accessing data.

1.3 Other Initiatives

The eBank UK project is of course not unique in offering metadata as a discovery tool for a data-rich repository. Examples from eScience initiatives include the NERC DataGrid, which developed an extensive metadata model, part of which was directed at discovery services [11].

The FEARLUS-G project designed an OWL ontology [12] that is used in a service that allows land-use scientists to access and re-use results and observations.

AstroGrid is part of the International Virtual Observatory Alliance collaboration, working towards interoperability of astronomical data. Standardisation of metadata has been one of their efforts, and a number of schema have been developed, including a recommendation on resource metadata.

The MyGrid Ontology supports a large number of components that deals not only with data but also workflows, people, specifications, organizations and services.

2. The eBank UK metadata application profile

The definition of a metadata profile to describe the datasets made available for discovery in this project has been an ongoing activity. Agreement has been reached gradually and the profile has evolved and been refined over time, through a number of iterations. The current product reflects the input of a multidisciplinary team with the digital library influence making a definite impact. Knowledge of the specific application area was provided directly by the crystallographers, all active practitioners in their field, who were involved throughout. Additionally, two workshops as well as desk research were used to gain a wider perspective of other scientific areas and to validate the outputs with community leaders.

2.1 Starting from Dublin Core

The metadata that is exposed from the dataset archive via OAI-PMH is intended to fulfil a dissemination role, leading to a number of discovery services. The Dublin Core standard has since its inception been intended as a resource discovery standard and carries support in the digital library communities. It was thus evaluated first and has now formed the basis of the metadata exposed by the crystallography metadata service.

The Dublin Core consists of an Abstract Model (of resources being described by metadata descriptions), a number of metadata terms used to describe resources, encoding guidelines (for XML and RDF), and schemas (in different schema languages) defining the Dublin Core terms. It should be noted that the abstract model was still in development when the eBank project started; it was finalised in March 2005. The eBank UK project has specified

- a number of Dublin Core metadata terms that can be used in the description of a crystallography data resource,
- their encoding in XML, and
- an XML schema definition for metadata exchange.

Furthermore, internationally-agreed guidelines exist for documenting the use of Dublin Core in a specific application (a so-called Application Profile) and documentation according to these guidelines is provided at [13]. The aim of this documentation is to assist others either wishing to create instances similar to the eBank repository, or to process the

metadata exposed by the repository, thus facilitating re-use.

2.2 Qualified Dublin Core

Dublin Core is presented as two levels. The simplest and most basic level consists of a basic element set. This is then extended to allow for element refinements, that is elements that refine (but not extend) semantics in ways that may be useful for resource discovery.

Two categories of refinements (also called qualifiers) are recognised; the first category makes an element narrower or more specific, e.g. the *relation* element can be refined to isVersionOf or isFormatOf. These are two specific instances of the relation element which narrow down the meaning. The second method of qualifying an element is by specifying an "encoding scheme" which aids in the interpretation of the element value. Examples of encoding schemes include controlled vocabularies or formal notations. Thus an element qualified by the encoding scheme will have as its value a token taken from a controlled vocabulary (e.g. a classification system) or a string formatted according to a formal notation e.g. expression of a date.

The eBank UK application profile makes use of the basic Dublin Core elements as well as the two mechanisms of refinements to define an element set that is specialized to the domain-specific needs. The 'dumbing-down' principle of Dublin Core ensures that applications that cannot recognise the specialised element can safely ignore the qualifier and treat the element as unqualified. Despite the loss of specifity this is intended to allow the remaining value of the element to be used correctly and usefully for resource discovery.

2.3 Generic Dublin Core descriptions

The Dublin Core metadata elements (indicated in this section by the prefix dc) comprise a number of elements that can (intentionally) be applied to a large number of different resources. For example dc:title is the name given to a resource. This label can be applied equally to the title of a book, a photograph, a work of art or an experiment. In the eBank application profile, the title of the datasets takes the name of the molecular compound that is being determined in the crystallographic experiment.

dc:date is a date associated with an event in the life cycle of a resource. Typically the date is associated with the creation or availability of the resource; the date was considered a useful element to include since it could allow users to limit searches for datasets added within a specific time period. A scientist (or service) could then generate a 'latest additions' or 'added since' feature.

dc:creator is the entity responsible for making the content of the resource. This has been interpreted as the names of the scientists themselves who create the datasets that are deposited. Previous work has identified issues of metadata quality and it is recommended that content rules should be applied to the values of metadata elements to improve quality [14]. One recommendation is that author names should be entered in controlled form and this has been implemented in the deposition software, so that name formats adhere to those recommended for eprints [15]. The organisations that the scientists belonged to was also included but this was designated within the dc:publisher element, which is defined as an entity responsible for making the resource available.

dc:type describes the nature or genre of the content of the resource. Various types of resources are described and disseminated using OAI-PMH (text, image, audio and video are but a few examples available from OAIster [16]), therefore the value of this element was considered important to enable selective and sorting operations on the harvested metadata in a heterogenous environment. The value currently implemented for the content of this element is 'crystal structure data holding'. There are other potential values that this element could take (not currently implemented), such as the more generic 'dataset'. Note that all elements in Dublin Core are repeatable therefore using all the values in repeated dc:type elements is an option.

dc:identifier is defined in the Dublin Core documentation as an unambiguous reference to the resource within a given context. In the context of the service being deployed in the eBank project, the most useful identifier would be one that not only identifies the datasets but can also be dereferenced to access the resource.

Two types of identifier are being used. The URL of the web entry for a holding in the archive is used as it provides an entry point to all the datasets for a crystal structure. On clicking the URL, the web page for the entry displays links that enable download of various datasets, and a selection of key metadata about the resource.

An alternative identifier is also being implemented. This is the Digital Object Identifier (DOI). DOIs are assigned in collaboration with a German agency. The DOI

system of identification includes a network of registration agencies. Alongside the management of identifier assignment, agencies also record metadata about resources registered. It is hoped that besides the technical details and infrastructure, this more formal approach to identifier assignment will inject a degree of commitment towards, and permanence of, the resources registered, thus ensuring persistence of identifiers and their resolution. This is a very relevant issue to harvesters of the metadata since unreliable identifiers would affect the quality of the discovery services provided using harvested metadata. The choice of DOI was partly influenced by collaboration with publishers in the crystallography literature, (described further in section 3.1), since the DOI has been especially successful in uptake with publishers.

dc:isReferencedBy is a refinement of dc:relation that is intended to contain pointers to published literature that specifically refer to the data in the archive. This element may contain either a textual citation to the publication, or an actionable URL, or other identifier.

dc:rights provides information about rights held in and over a resource. This is crucial information both for the data provider who will want to assert the rights held and granted over a resource that is being made available, and dually to a potential user who has discovered the resource. Discussion with third parties planning to use the harvested metadata indicated that this information would be absolutely necessary for them to determine what sort of access and use of the resources was allowed, thus determining what discovery services could be provided. In the eBank implementation the rights are described by means of a statement declared on a web page, the URL for the page is then exposed in the dc:rights metadata element. This is possible since all the resources in the repository have been given the same right of use by the depositors.

2.4 Chemistry-based descriptions

Crystallographic determinations involve the analysis of chemical structure and the resulting datasets often consist of three-dimensional co-ordinates. However, for the purposes of naming the structure, more condensed forms (textual and/or formulaic) are used. These can all be of assistance to the crystallography researcher in identifying molecules of interest and are thus important metadata to include in a search service.

One example of how chemical names have been included in the eBank data is the use of the International Union of Pure and Applied Chemistry (IUPAC) name in the title. The generation of this title requires the use of guidelines and the contribution of human expertise as well as software processing.

Additionally, a number of other chemistry metadata is provided in multiple dc:subject metadata elements. dc:Subject describes the topic of a resource and since the chemical structure is the main topic of the datasets, the different ways of identifying the structure have been included. Since the different naming mechanisms have different rules for how they can be formulated and formatted, a typical search service would need knowledge of the formats present in the metadata, so that appropriate search criteria can be given to the user. The Dublin Core encoding schemes refinement mechanism has been applied to distinguish between the different naming conventions. Thus chemical formulae are considered to be taken from the controlled vocabulary that consists of all chemical formulae that can be expressed. Similarly, InChIs (a recently agreed international standard for uniquely identifying molecules) are considered to be another controlled vocabulary which can be applied to dc:subject to restrict the values which that element can take when this encoding scheme is declared.

Another form of controlled vocabulary has also been defined by the eBank project. This consists of a small set of terms which assign the dataset to one of four types of compound class (organic, inorganic, bio-organic, organometallic). Once again this is a specialised type of vocabulary that was considered useful in the application domain to enable users to narrow down searches to their field of interest when carrying out discovery activities.

One further use of the encoding scheme mechanism provided by Dublin Core was made to define a list of types for datasets. The values of dc:type in eBank can be further refined by declaring the type to be one of the eBankDatasetTypes. These consist of a closed list (though extension in the future is possible) and are declared in the published XML schemas.

2.5 Exposing Complexity

One aspect of the repository that has been glossed over so far is the granularity of the resources provided. Each crystal structure determination consists of a number of experimental stages, with datasets being produced at each stage. The datasets may vary,

from images of x-ray diffraction patterns, to structured documents, such as the CIF. One aim of the repository was to improve on the current publication and dissemination processes (which make the CIF file available) by giving access to **all** the results from a determination. Each entry in the repository therefore consists of a collection of files from different stages of the experiment, all arising from the determination of one chemical structure.

From the discovery perspective, the metadata should reveal to the user the existence of the multiple files available, together with an indication of their nature. In this specific domain, the experimental stage from which the files are generated is considered significant and indicative of the files available. Altogether, the files make up a collection, or more specifically a data holding, and share common characteristics such as the creator, and chemical descriptors.

The model of the crystallography resource that emerges bears some resemblance to a category of digital objects termed *complex objects*. In the digital library domain, content packaging standards are being proposed as a tool to disseminate digital resources that are composed of more than one underlying file or object. Typically these standards consist of a mechanism for declaring the structure and make-up of digital items, and use XML schemas that define an XML format that either references or contains the data files that make up the package.

METS [17] is one such packaging format being investigated by the eBank UK project. METS is maintained by the Library of Congress; a METS document consists of seven major sections, including a METS header and descriptive metadata. The descriptive metadata section is designed so that metadata descriptions (e.g. those in Dublin Core) can either be included or referenced. This allowed the project to re-utilise the Dublin Core descriptions, outlined above, within a METS file.

The File section and the Structural Map section of METS are used to link (or contain) the files, and to describe the logical or physical relations, respectively. These are the sections currently being implemented and investigated.

The Behaviour section of METS (in common with other sections of some of the other packaging formats) associates files with executable code needed to read or manipulate them. The eBank project has so far concentrated on the dissemination role of the metadata, simply advertising the existence of datasets; specialised sources already exist with integrated facilities for access coupled with data manipulation, and it was not the intention of eBank to compete with these sources. We therefore do not have any immediate plans to implement this feature of the METS standard. However it is an obvious advantage that the standard accommodates this information which is available for others to use should they wish to extend on the eBank output.

3. The Realities of Cross-Discovery.

As a cross-disciplinary effort, one of the interests of the eBank UK project was to encourage and explore common technologies across disciplines and resource types, with a focus on linking between data sets and published literature. Given its uptake in the digital library world, and the experience of the project partners, the OAI-PMH was viewed as a promising candidate protocol to provide shared technology connecting digital libraries with scientific repositories.

3.1 Connections with published literature

Initially it was envisaged that an OAI-PMH repository of crystallography literature with known connections to the deposited data would be created for initial demonstration of cross-search capabilities. Unlike some other disciplines, (such as physics), there was no existing body of publications of crystallography that was already accessible in this way. This plan was superceded since good collaborations that developed between project partners and the IUCr resulted in publication metadata being made available by the latter body for the purposes of demonstration in a web interface. Thus XML descriptions of a small sample of carefully selected publications were cross-searched with the metadata from the data archive.

The searching capabilities were shown in a demonstrator prototype which supported searching against author/creator names and chemical information. The publications included had been identified such that connections were known to exist between the articles and the datasets.

The use of the OAI-PMH had an unplanned consequence. Discussions are in an advanced stage to allow publishers to use the protocol to access data that is usually submitted in support of published articles. The use of the DOI also means that rather than reproduce the data in the published article (as sometimes happens), the data will simply be referenced with more space in the article devoted to discussion of the

results. Thus although not exactly as originally intended, OAI-PMH looks set to become part of the infrastructure for publishing data in crystallography, and this may lead to new dissemination routes for the data.

3.2 Connections with eLearning

One other angle that the eBank project is exploring is the pedagogical role of data. A study is currently being undertaken to examine the interaction of students directly through the web interface of the archive containing the data. However, from the discovery perspective, the potential of the metadata to be used to identify relationships or relevance between datasets and e-learning materials from other sources is still largely unknown.

The OAI-PMH is being used in a growing repository of learning resources funded by the JISC. JORUM [18] is a free online service for teaching and support staff in UK Further and Higher Education Institutions, helping to build a community for the sharing, reuse and repurposing of learning and teaching materials. It is hoped that searches can identify resources from JORUM that complement the datasets in the eBank repository, when used by students to search for data.

With such a specialist area as that addressed by eBank to date, it is always questionable whether the coverage of a generic resource such as JORUM will be wide enough to contain a sample of results relevant to an eBank search. Initial examination of the metadata revealed that a potential subset of resources (albeit small in number) were relevant to the topic of crystallography. Regretfully, at the time of writing access to the learning resources themselves was still being negotiated (institutional sign-up and ATHENS authentication is required). Therefore their actual relevance and the target audience could not be evaluated by the crystallography users in a demonstrator prototype. It is hoped that once the access issues are sorted out a demonstration and evaluation can take place.

3.3 Validating the output

Ideally, the true test of the metadata profile would be demonstrated bv a cross-search service against a number of different repositories containing either crystallographic, or indeed other resources. Due to the lack of suitable compatible repositories (i.e. ones supporting OAI-PMH) it is still too early to be able to carry out such validation.

3.4 Use of Dublin Core and OAI-PMH in eScience

[19] in a report on data curation for eScience in the UK in 2003 found an “almost total lack of knowledge of metadata tools such as the Dublin Core” in the responses to a questionnaire. However eScience initiatives (such as those quoted in Section 1.4) do report using some of the Dublin Core elements amongst a number of other specialized descriptions adapted to their applications. There is not much evidence however of use of the OAI-PMH as a common framework to deliver research data or metadata (at least in the UK) to date.

4. Future Work

The eBank UK project has defined the metadata fields (described in sections 2.3 and 2.4) based on typical laboratory practice at a large national centre, with the aim of providing an example that could be re-used in other settings. Furthermore, the software that generates the metadata is specialised to the workflow at that laboratory. Work is ongoing at other centres to evaluate alternative OAI-PMH software in different laboratories, and it is hoped to evaluate the eBank software modifications against different laboratory practices in the future.

So far only a relatively small sample of datasets have become accessible and searchable through the eBank initiative. It is therefore difficult to evaluate the efficiency of searching the metadata fields exposed until a critical mass of metadata is available. Furthermore, for some of the fields used (e.g. the InChI), experience is still being gained within the discipline to establish helpful searching methods (e.g. substructure searching).

However the notion of open access for crystallography data has been widely disseminated within the crystallography community, and generally very well received. Discussions are ongoing and funding is being sought to enable co-operation with other existing repositories to expose their metadata using OAI-PMH. These efforts together with existing activities may in the future allow for a more extensive evaluation of cross-searching and related discovery issues, with feedback from a larger group of users, beyond proof of concept.

The inclusion of a DOI as a citation for a dataset when referenced in a publication is still in the very early stages (although at least some publishers are receptive to the idea). This practice will be tested in the near future so that the reaction and interest of the community can

be gauged. Other quantitative measures of success will include actual access and download of the datasets as a result of the links appearing in the literature.

The present architecture of the eBank UK project centres on the OAI-PMH. However it is relatively straightforward to support other protocols for searching the harvested metadata, for example the demonstrator search service uses Z39.50 software to support its indexing and searching functionality. It would be possible to expose the metadata to outside parties using that protocol or related ones (such as the SRU [20]) if the use case was made.

5. Conclusion

As increasing amounts of data are generated and made available for sharing and re-use through eScience initiatives, the curation responsibility of enabling discovery of such data will become a more pressing issue. The eBank UK project has demonstrated the use of the OAI-PMH as a technology for metadata-mediated discovery in crystallography. For this protocol to become the connecting technology between repositories of crystallography data, and provide a shared infrastructure across different resource types, a more widespread adoption of the protocol amongst the applicable repositories would be required.

6. Acknowledgment

The eBank UK project http://www.ukoln.ac.uk/projects/ebank-uk/ is funded under the JISC Semantic Grid and Autonomous Computing Programme. The project partners are UKOLN, University of Bath, School of Electronics and Computer Science, University of Southampton, The School of Chemistry, University of Southampton and PSIgate, University of Manchester. The work reported here is a result of collaborative effort between project members, (past and present) from these institutions.

7. References

[1] Rusbridge C., et al *The Digital Curation Centre: A vision for Digital Curation.* From Local to Global: Data Interoperability — Challenges and Technologies, Mass Storage and Systems Technology Committee of the IEEE Computer Society, 20–24 June 2005, Sardinia, Italy

[2] Mc Mahon, B. *Semantically Rich Metadata in Crystallographic Publishing.* EUNIS 2005, University of Manchester, UK.

[3] International Union of Crystallography. http://www.iucr.org/

[4] Crystallography Open Database. http://www.crystallography.net/

[5] RCSB Protein Data Bank. http://www.rcsb.org/pdb

[6] The Open Archive Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/

[7] Duke, M., Day, M. ,Heery, R. et al. *Enhancing access to research data: the challenge of crystallography.* In: Proceedings of the 5th ACM/IEEE Joint Conference on Digital Libraries, Denver, CO., USA, June 7-11, 2005. New York: Association for Computing Machinery, 2005, pp. 46-55. ISBN 1-58113-876-8.

[8] Heery, R., Duke, M., Day, M. et al. *Integrating research data into the publication workflow: eBank experience* In: Proceedings PV-2004: Ensuring the Long-Term Preservation and Adding Value to the Scientific and Technical Data, 5-7 October 2004, ESA/ESRIN, Frascati, Italy, Noordwijk: European Space Agency, 2004, pp. 135-142.

[9] Coles, S., Frey, J., Hursthouse, M. et al. *Enabling the reusability of scientific data: Experiences with designing an open access infrastructure for sharing datasets.* Designing for Usability in e-Science. International Workshop, NeSC, Edinburgh, Scotland, 26-27 January, 2006

[10] Coles, S.J., Frey, J.G., Hursthouse, M.B. et al. *An e-Science environment for service crystallography - from submission to dissemination.* Journal of Chemical Information and Modeling, Special Issue on eScience. 2006 (In Press).

[11] O'Neill, K., Cramer, R., Gutierrez, M., et al. *A specialised metadata approach to discovery and use of data in the NERC DataGrid.* Proceedings of the U.K. e-science All Hands Meeting, 2004.

[12] Pignotti, E., Edwards, P., Preece, A., et al. *Providing Ontology Support for Social*

Simulation. Proceedings of the First International Conference on eSocial Science, NCeSS/ESRC, Manchester, 2005.

[13] The application profile for crystallography data. http://www.ukoln.ac.uk/projects/ebank/schemas/profile

[14] Guy, M., Powell, A and Day, M. *Improving the Quality of Metadata in Eprint Archives* Ariadne, Issue 38, January 2004. http://www.ariadne.ac.uk/issue38/guy/

[15] Powell, A., Day, M and Cliff, P. *Using simple Dublin Core to describe eprints.*, March 2003 http://www.rdn.ac.uk/projects/eprints-uk/docs/simpledc-guidelines/

[16] OAIster. http://oaister.umdl.umich.edu/

[17] Metadata Encoding and Transmission Standard. http://www.loc.gov/standards/mets/

[18] JORUM. http:///www.jorum.ac.uk

[19] Lord, P. and McDonald, A. eScience Curation Report, JISC, 2003

[20] Search/Retrieval via URL http://www.loc.gov/standards/sru/