Facilitating the Deposit of Experimental Chemistry Data in Institutional Repositories: Project SPECTRa (Submission, Preservation, and Exposure of Chemistry Teaching and Research Data)

Peter Morgan
Cambridge University Library, United Kingdom
pbm2@cam.ac.uk

Abstract

Institutional Open Access repositories are becoming established as an important part of the university library and information services infrastructure. While early efforts to populate them with content have concentrated on the deposit of peer-reviewed research papers, there is a growing awareness of their potential as repositories of data and other non-text materials, and consequently a need to develop strategies and procedures that can realise this potential.

Chemistry as a discipline has been slower than the physical and biomedical sciences to adopt and exploit Open Access concepts in the handling of experimental data and research publications. Chemical information is essential to many sciences outside chemistry, and the reporting of the synthesis and properties of new chemical compounds is central to this. But most of the essential experimental data associated with peer-reviewed publications from chemistry departments are never communicated to the scientific community. These data are all available in high-quality electronic form in the laboratories but there is no effective method for archiving them or making them openly accessible.

The SPECTRa (Submission, Preservation, and Exposure of Chemistry Teaching and Research Data) project addressed this problem. It was a JISC-funded 18-month collaboration, ending in March 2007, between the university libraries and chemistry departments of the University of Cambridge and Imperial College London, in co-operation with the eBank-UK project. Its main objective was to develop a set of customized software tools that would enable chemists routinely to deposit experimental data in Open Access repositories, employing the DSpace repository platform used by the two libraries. The work was informed by surveys of research chemists in the two universities, exploring their use of information technology and assessing their interest in using repositories and Open Access principles for data management.

This paper presents the project's outcomes and discusses the implications for the development of library-managed institutional repositories.

Keywords: Chemistry Data, Data Repositories, DSpace, Institutional Repositories, Libraries, Open Access

Introduction

Although initial efforts to define and promote Open Access (OA) concentrated on the role of peer-reviewed research literature, the 2003 Berlin Declaration on Open Access [1] extended the vision with its statement that "Open Access contributions include original scientific research results, raw data and metadata, ...[etc.]"; and the OECD followed suit in its 2004 Declaration on Access to Research Data from Public Funding [9] when it recognised the value of placing scientific research data in openly accessible data

collections. Hey and Trefethen [4], discussing the UK's e-Science programme, described the "vast outpouring of scientific data" and noted the need "to automate the discovery process - from data to information to knowledge - as far as possible." The concept of Open Data [9] has evolved alongside, but is not synonymous with, Open Access, since a scientific research paper may be available through OA while the associated data remain locked away behind access restrictions.

One conspicuous component of the OA movement has been the development of repositories as a means of managing the deposit, dissemination and preservation of research outputs in digital form. The Directory of Open Access Repositories, OpenDOAR [11], currently (April 2007) lists 855 repositories worldwide, of which 80% have been established as institutional repositories, usually managed by libraries; but while text-based materials dominate the types of content found in all repositories, only 6% contain datasets.

Despite the extensive efforts made by advocates of OA internationally and locally, these repositories are acquiring content only slowly, and many academics - however supportive in principle - remain reluctant in practice to deposit their digital content in any available repository. Faced with these difficulties, institutional repository managers are increasingly seeking to analyse the reasons for such reluctance and identify obstacles to wider compliance with the aim of refining their organisational strategies and developing new procedures that will encourage researchers to utilise institutional repositories as a routine part of the research process.

The institutional background

The University of Cambridge and Imperial College London are both research-intensive universities, and both are consistently ranked among the top three universities in the UK, with 87 Nobel Laureates having been affiliated to one or other of the two institutions. Ensuring that their research outputs can be disseminated and preserved is thus a high priority in determining their institutional repository strategies.

Cambridge University Library established its institutional repository, DSpace@Cambridge [2], in 2003, initially as a project and since 2006 as a formal service run in collaboration with the University Computing Service. From the outset its institutional repository policy avoided imposing restrictions on the types of content it would accept, preferring instead to develop a dialogue with the University's researchers across all disciplines in order to identify how best the institutional repository could meet their needs. As a result of this approach significant collections of images, video, and data were acquired. In particular, researchers in the Chemistry Department's Unilever Centre for Molecular Informatics, led by Dr Peter Murray-Rust, formed one of the repository's most enthusiastic 'early adopter' communities, depositing a large quantity of OA data files describing molecular structures.

However, the Library's experience with the chemists also highlighted the difficulties of moving beyond a core group of OA and institutional repository enthusiasts and gaining acceptance among the mainstream body of chemistry researchers. It became clear that many of the established practices involved in archiving and publishing experimental data from chemistry research were a major obstacle to widespread adoption of the institutional repository, leading the Library and the Unilever Centre chemists to conclude that further work on deposit procedures, customising them to meet the needs of the chemistry research community, was a highly desirable next stage if a suitable opportunity could be identified.

The search for potential project partners led immediately to Imperial College London. Many of Murray-Rust's research interests were shared by Professor Henry Rzepa in the Computational Chemistry Department at Imperial: long-standing research collaborations between the two had resulted, among other things, in their development of Chemical Markup Language (CML) [8]. At the same time Imperial College Library was developing

its own plans for a DSpace-based institutional repository that would include both text materials and supporting datasets across a number of scientific disciplines.

The SPECTRa Project [12]

In 2005 Cambridge University Library was awarded an 18-month grant from the UK Joint Information Systems Committee (JISC)'s Digital Repositories Programme [6] to fund the SPECTRa Project as a partnership between the University of Cambridge and Imperial College. In addition to the formal project partners, SPECTRa also entered into an agreement on collaboration with another longer-established JISC-funded project, eBank-UK [3], which was exploring ways of integrating crystallographic datasets into digital repositories.

The SPECTRa project team consisted of three directly-employed personnel (project manager, software developer, and [at Imperial] project officer) together with senior staff from the two libraries and chemistry departments. In Cambridge we decided, largely because of the distance (c. 2km) from the Library, that the project personnel would be based in the Chemistry Department in order to be close to the target research community; whereas at Imperial, where the Library and the Chemistry Department are very close, the Library was used as the project base. The work of the team was overseen by a Steering Group comprising senior members of both partner institutions together with representatives from eBank-UK and external members.

The SPECTRa project plan took as its starting point the proposition that chemistry as a discipline has been slower than the physical and biomedical sciences to adopt and exploit Open Access concepts in the handling of experimental data and research publications. At least 80% of data (analytical, spectral and even crystallographic) associated with peer-reviewed publications from chemistry departments are never communicated to the scientific community. In those limited instances where a publisher does provide a means of accessing primary data to supplement a published paper, the data may then be subject to the publisher's IPR practices, and in most cases the primary data are simply not published. For example, chemical theses contain spectra that are not routinely captured and exposed to search tools, and that are typically stored without being subjected to appropriate preservation techniques, with the likely irretrievable loss of data within a few years. [14]

The project's primary aim was to investigate the needs of the academic chemistry research community in capturing and re-using experimental scientific data, and to facilitate the routine extraction of data in high volumes and their ingest into institutional repositories. To achieve our objectives we set out to:

- investigate the needs of the academic chemistry research community for the capture, long-term storage and re-use of experimental data in digital repositories;
- demonstrate how these needs may best be co-ordinated with emerging institutional strategies for repositories handling both data and publications;
- interview researchers about their requirements for a repository toolset which will facilitate routine extraction of data in high volumes and ingest these data into institutional repositories as part of their normal workflow;
- develop context-specific metadata based on Dublin Core and on work already developed by eBank-UK;
- investigate the cultural issues in capturing and re-using scientific data, including the willingness of researchers to submit their work to Open Access publication; and
- explore interoperability issues involved in archiving data in repositories.

Methodology

The project had two key elements: surveys of researchers in order to identify their expectations, needs, and problems; and the design and development of software tools to

facilitate data deposit. The emphasis was on creating a repository architecture for data submission and preservation, with practical tools that could facilitate the process, thus providing chemists with a user-friendly system that would encourage them to deposit their experimental data routinely.

Investigating the needs of the chemistry research community

The project selected three distinct areas of chemistry research – synthetic organic chemistry, crystallography and computational chemistry - for investigation. Each of these proved to have specific requirements, reinforcing the belief that institutional repositories needed to be aware of such issues and to respond by developing discipline-specific procedures.

We originally anticipated that we would need to develop detailed protocols for chemists' workflows, but preliminary investigations showed that these were not necessary in synthetic and computational chemistry: in these disciplines, only the data produced as the end-point of the experimental or calculation process were regarded by chemists as having sufficient value to merit archiving.

For synthetic chemistry, the project team initially undertook interviews with researcher leaders. The results indicated that they were markedly reluctant to use a repository unless unpublished or commercially-sensitive data could be protected by an embargo procedure. This early finding became a central focus for the project's subsequent work on data management policy and practice.

In addition to the individual interviews, we carried out a broader survey of all research chemists (postgraduate students, postdoctoral workers and academic staff) at both Imperial College London and the University of Cambridge, using a 28-part questionnaire. Its purpose was to assess how respondents used computers and the Internet, and to identify specific data-handling practices and needs. Participation was voluntary, and we obtained an overall response rate of 22%. Follow-up interviews were conducted with some respondents. A detailed report on the survey is available as an appendix to the SPECTRa Project final report [13]. Its major findings were:

- much data (e.g. lab books, paper copies of spectra) is not stored electronically;
- a complex list of data file formats (particularly proprietary binary formats) is being used;
- there is a significant ignorance of digital repositories among chemistry researchers;
- a requirement was identified for experimental data in repositories to be available only on restricted access; and
- the ability to search the repository by chemical *substructure* (a level of granularity not supported in current metadata practice) was seen as the most essential facility for a working system.

Neither crystallography nor computational chemistry lent themselves to a similar survey approach. In the former case, interviews with departmental crystallographers - a small and specialised service group providing confirmation of new chemical structures - gave us a basis for understanding their data-archiving requirements; while the scale and complexity of computational chemistry methods was such that the project team focussed their efforts on one aspect only, that of Gaussian calculations.

Software tools

Wherever possible, the software development programme reused existing Open Source code, and from the outset it was understood that all software newly created in the course of the project would also be made available as Open Source code. Linked to the browser-based client tools for file uploading, the SPECTRa deposit tools interfaced with processes

for automated CML creation, file validation against available specifications, metadata extraction, and METS [7] packaging. (METS - Metadata Encoding & Transmission Standard - was adopted as the most appropriate schema, both because it was a relatively simple technology and because DSpace already supported it.) File validation could pose particular problems as proprietary file formats are widely used in chemistry and use non-open standards. Chemical metadata fields would routinely include systematic name and the unique InChI (International Chemical Identifier) [5] which provides an exact means of chemical structure searching.

Repository platform

As both partner institutions were already committed to using DSpace as the technology platform for their institutional repositories and - particularly at Cambridge - had acquired considerable experience in running and developing it, the decision to use the same platform for SPECTRa was a logical step. As the project deliverables did not include actual deposit of data into the institutional repository, we agreed that work on the DSpace technology would concentrate on installing and developing local instances of the DSpace software for use by the two chemistry departments. This allowed us to explore issues arising from the relationship between departmental and central institutional repository installations.

Summary of findings

We set out to study how institutional repositories might develop a closer relationship with researchers, using chemistry as the area for investigation, and to test whether repositories offered an acceptable way of managing experimental data. Our main conclusions, from a repository management point of view, may be summarised as follows.

- Repositories handling scientific research outputs are complex to build and maintain, and need to be responsive to the specific requirements of individual disciplines.
- Purpose-built deposit tools can be used to encourage researchers to deposit their data, provided that these tools are compatible with their established workflows.
- There is a notional "golden moment" for data capture: this is the point at which the researcher is most likely to understand the process, be in possession of a comprehensive package of information to describe it, and have the motivation to deposit it into a data management system such as a repository.
- To meet researchers' concerns about the deposition of unpublished or commercially-sensitive data in a repository, an embargo process is required to ensure that data files will be made openly accessible only at a date agreed by the researcher. The SPECTRa architecture envisages that all such data will initially be deposited in a designated 'embargo repository'. At that point the researcher is actively required to specify the length of the embargo and whether, at the end of that period, the data should automatically be released to an OA repository or reviewed. The embargo information is held in the metadata.
- Software tools can automate data validation and metadata creation to a substantial extent, but the deposit process will require a degree of editorial intervention to manage certain tasks (e.g. author name identification).
- Persistent identifiers are essential for the effective long-term management of files in repositories. Publishers favour the DOI (Digital Object Identifier) system, which requires an annual renewal fee for each item and thus has serious long-term cost implications where large numbers of data files are involved.
- DSpace employs the Handle system for its persistent identifiers. Its functionality does not yet permit Handles to be updated when files are moved from one repository to another. Until this is more fully developed, institutional repositories based on DSpace will not be able to implement all aspects of the distributed departmental-central architecture envisaged by SPECTRa.

- For all data being submitted to a repository, ownership and licensing arrangements for data re-use need clear guidelines that can be applied consistently across multiple institutions. This is an essential prerequisite for the release of data in accordance with Open Data principles.
- Science is driven by data, and quality data must be valued as a major asset. Institutions and research funders should therefore be encouraged to recognise this and allocate appropriate resources for data management, dissemination and preservation.
- The architecture of an 'institutional' repository may embrace a co-ordinated network of departmental repositories. These would be responsible for capturing and managing files at a point close to the researchers, with content subsequently deposited into the central repository for large-scale aggregation and long-term preservation. (Fig.1)

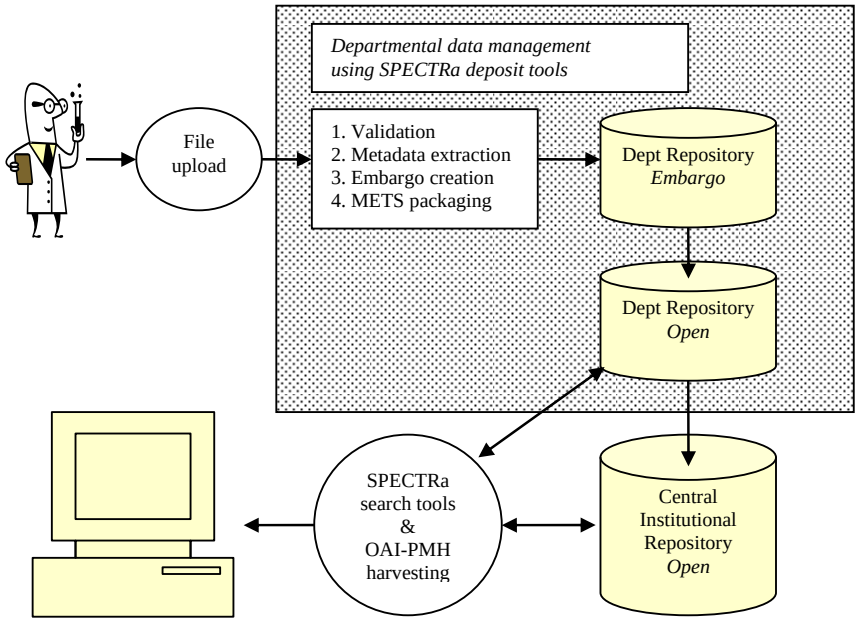

Fig.1 Overview of SPECTRa Project architecture

Conclusions

SPECTRa has demonstrated the benefits of working closely with researchers, but to do so is a very resource-intensive process. It is unlikely that most institutional repository managers will have either the time or the specialised expertise to engage in such narrowly-focussed work across all disciplines within a single institution. Progress is thus more likely if generic solutions can be pursued and shared for implementation across multiple institutions, although it may be difficult to achieve a satisfactorily balanced compromise between the specific and the generic.

The SPECTRa repository architecture has assumed that the central institutional repository will normally be the appropriate location for long-term preservation of data files, and that departmental repositories would act as holding areas for short- or medium-term file management. However, the decision as to whether departmental repository content is eventually transferred to the institutional repository is ultimately a matter be decided by local policy. Policy decisions about long-term retention are subject to stakeholder requirements at both local and national levels, and were outside the scope of the project.

While the concept of a central institutional repository offering a managed dissemination and preservation service remains valid, there are both organisational and technical arguments in favour of creating departmental repositories that offer services customised to the needs of the departmental research community and feed content to the institutional repository, all within the overall architecture of the institutional repository framework The implications of this approach for institutional repository management and policy development are considerable.

Acknowledgements

The author is indebted to his colleagues on the SPECTRa Project Team, without whom this paper would not have been possible: Jim Downing, Dr Peter Murray-Rust, Dr Alan Tonge (all University of Cambridge); Fiona Cotterill, Janet Evans, Professor Henry Rzepa, Lorraine Windsor (all Imperial College London).

The SPECTRa Project Team are grateful to the Joint Information Systems Committee (JISC) for funding this work, and to the JISC Digital Repositories Programme staff for their guidance and support.

References

[1] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (October 2003) http://oa.mpg.de/openaccess-berlin/berlindeclaration.html *[accessed 23 April 2007]*

[2] DSpace@Cambridge: http://www.dspace.cam.ac.uk/ *[accessed 23 April 2007]*

[3] eBank-UK Project: http://www.ukoln.ac.uk/projects/ebank-uk/ *[accessed 23 April 2007]*

[4] Hey, A.J.G. & Trefethen, A.E. (2003): 'The Data Deluge: an E-Science Perspective'. In Berman, F., Fox, G. C. & Hey, A. J. G., Eds. *Grid Computing - Making the Global Infrastructure a Reality*, chapter 36, pp. 809-824. Wiley and Sons. http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/research/esci/datadeluge.pdf *[accessed 23 April 2007]*

[5] InChI (International Chemical Identifier): http://www.iupac.org/inchi/ *[accessed 23 April 2007]*

[6] JISC Digital Repositories Programme http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories.aspx *[accessed 23 April 2007]*

[7] METS (Metadata Encoding & Transmission Standard) http://www.loc.gov/standards/mets/ *[accessed 23 April 2007]*

[8] Murray-Rust, P. & Rzepa, H.S. (2003): "Chemical Markup, XML and the Worldwide Web. 4. CML Schema", *J. Chem. Inf. Comp. Sci.*, **43**, 757-772 DOI:10.1021/ci0256541 *[accessed 23 April 2007]*

[9[OECD Declaration on Access to Research Data from Public Funding (January 2004) http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html [Annex 1] *[accessed 23 April 2007]*

[10] Open Data: http://en.wikipedia.org/wiki/Open_data *[accessed 23 April 2007]*

[11] OpenDOAR Directory of Open Access Repositories http://www.opendoar.org/ *[accessed 23 April 2007]*

[12] SPECTRa Project: http://www.lib.cam.ac.uk/spectra/ *[accessed 23 April 2007]*

[13] SPECTRa Project - Final Report: http://www.lib.cam.ac.uk/spectra/FinalReport.html *[accessed 23 April 2007]*

[14] SPECTRa-T, (http://www.lib.cam.ac.uk/spectra-t/), a new JISC-funded project involving the same partner institutions as SPECTRa, will address the extraction of experimental chemistry data from e-theses using text-mining tools.