

SPECTRa

Submission, Preservation and Exposure of Chemistry Teaching and Research Data

A proposal submitted by
the University of Cambridge (lead partner) and Imperial College London
to the JISC Digital Repositories Programme (call 03/05)

A. Introduction

- 1.1 This project will meet objectives in activity areas (ii) and especially (iii) of the JISC call for projects in digital repositories. It will address the provision of Open Access to primary chemical research data in molecular and related subjects through the use of institutional repositories. It will build on the work of eBank UK, collaborating closely with members of that project to ensure optimal harmonisation of effort and outcomes. It will also build on experience already acquired by the Chemistry departments at Cambridge University and Imperial College London in Open Access publishing of scientific data, and will integrate this with the same universities' institutional strategies and practices for library-managed OAI-compliant repositories, thus creating a testbed of relevant scientific resources and expertise. The project will initially study the needs of researchers and scope their data-handling needs. It will use the findings of this study to develop automated tools so that high-volume data can be identified, extracted, and archived in repositories, where it will be preserved and accessible for use to support research and teaching. The study's outcomes will be analysed to provide guidance of relevance to the JISC's chemistry research community, and its methodology will be formulated to provide generic guidance for similar studies in other sciences. The tools subsequently developed will be available as Open Source code designed for use with the DSpace repository platform.
- 1.2 The project will last for eighteen months (June 2005 -November 2006).
- 1.3 The project will contribute to Activity Area (ii) through its study of the workflow practices and needs of research chemists. This study will identify the data types requiring more effective archival management and enable the project team to design automated tools for archiving that can be embedded in workflows. It will contribute to Activity Area (iii) by creating, testing, and implementing production tools that provide an automated interface between the creation of experimental data in chemistry research and the archiving of the same data in institutional repositories.

Aims

- 1.4 The aims of SPECTRa are to:
 - investigate the needs of the academic chemistry research community with respect to how data associated with theses and peer-reviewed publications may best be communicated.
 - demonstrate how these needs may best be co-ordinated with emerging institutional strategies for repositories handling both data and publications.
 - facilitate routine extraction of data in high volumes and their ingest into institutional repositories.
 - investigate the cultural issues in capturing and re-using scientific data.
 - explore interoperability issues involving archiving data in repositories.

Objectives

- 1.5 SPECTRa will realise these aims through the following objectives:
 - undertake surveys of communities in computational and organic chemistry.
 - test and refine crystallography tools developed by eBank.
 - develop automated validated and indexing tools specific to computational chemistry and organic spectra, and providing interactions with the DSpace repository platform.
 - develop chemical metadata functionality based on Dublin Core.
 - disseminate and promote project outcomes to encourage widespread adoption.

B. Project description

- 2.1 The 2003 Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities states that Open Access contributions will include "original scientific research results, raw data and metadata, source materials, digital representations of pictorial and graphical materials and scholarly multimedia material".¹ More recently the 2004 OECD Declaration on Access to Research Data from Public Funding has recognised the value of placing scientific data in openly accessible data collections.² Hey and Trefethen, discussing the "vast outpouring of scientific data", have noted the need "to automate the discovery process - from data to information to knowledge - as far as possible."³
- 2.2 This proposal will address these issues, relating specifically to experimental data in chemistry. To complement other initiatives investigating central, national, and international subject-based repositories, it will explore ways in which institutional repositories can capture locally-generated data and contribute to a shared resource across a distributed grid of networked repositories. It will establish generic tools, general protocols and standards to ensure consistency of practice across different repository platforms.
- 2.3 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. Most of the 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. 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
- 2.4 But chemical information is essential to many sciences outside chemistry, including materials, life sciences and environmental and supports major industries including pharmaceuticals. The reporting of the synthesis and properties of new chemical compounds is central to this, with over 500,000 new syntheses in peer-reviewed publications from the global academic sector. The bare essentials of the synthesis are published but the essential experimental data are almost always omitted. Moreover the text-based nature of traditional publishing makes it extremely expensive to add chemical metadata to publications (as is done by Chemical Abstracts).
- 2.5 It has been reported⁴ that 80% of all crystallographic data are never published and we estimate that in organic chemistry 99% of all spectra (which are essential for the full quality control and understanding of chemistry) are lost. These data are all available in high-quality electronic form in the academic laboratories but there is no effective method for archiving them (the most favourable is that departing graduate students hand over floppies to supervisors, and this information rapidly decays). Most of those intrinsically high-quality objects decay with a short half-life and may be irretrievable within five years of their creation.
- 2.6 OAI-compliant institutional repositories are potentially an effective means of capturing, preserving, and disseminating them in accordance with Open Access principles. Supported by the UK eScience programme we have created proof-of-concept for the extraction of this data from the scientists and its archiving in repositories (EPrints/Southampton⁵ and DSpace/Cambridge⁶). Collections of 10,000 compounds could be managed in a research group and millions can be centrally archived. By adding chemical metadata⁷ (e.g. the new IUPAC unique identifier, InChI⁸) we can get essentially 100% precision and 100% recall from web-based search engines (Google, MSN, etc.) which harvest our repositories.
- 2.7 Methods for depositing much of this eChemistry are largely solved in principle but not deployed in HE departments. This project will produce a demonstrator of the cultural and technical advantages of deposition. The challenge is to show senior scientists in academia and publishing houses of the value and cost-effectiveness of information capture and archiving. The components (Figure 1) are:
 - 2.7.1 Streamlining and porting of the current technology. Our aim will be to develop a portable, free/Open, desktop tool into which scientists can dump their data. The tool will extract and systematise the types of data and extract and formalize (RDF, RSS) the chemical metadata. The data will then be automatically deposited in a central holding repository. We propose three main inputs of information (in increasing order of difficulty):

- crystallography (Cambridge and Imperial). We collaborate closely with the Southampton crystallographers on ePublishing. They have developed software and protocols as part of the eBank programme⁶ which they use for populating eCrystals/ePrints directly from experimental laboratories. We provided them with InChI/CML technology available, and jointly showed its value as metadata.⁵ This technology is now proven and mature and we will reimport it to show that it can be used for DSpace.
- computational chemistry (Imperial). We estimate that 100,000 - 1 million calculations of compounds and their properties are made annually and none are effectively deposited. We have developed automatic transmission, including chemical metadata to DSpace and subsequent indexing by search engines.
- organic synthesis (Cambridge). Many theses and papers consist largely of text-based reports of chemical synthesis. We will investigate the effort required to archive current electronic chemical information (normally in proprietary form) into XML/RDF+metadata in repositories. This builds on several years' work with our collaborating publishers and although the least predictable component will have the highest impact.

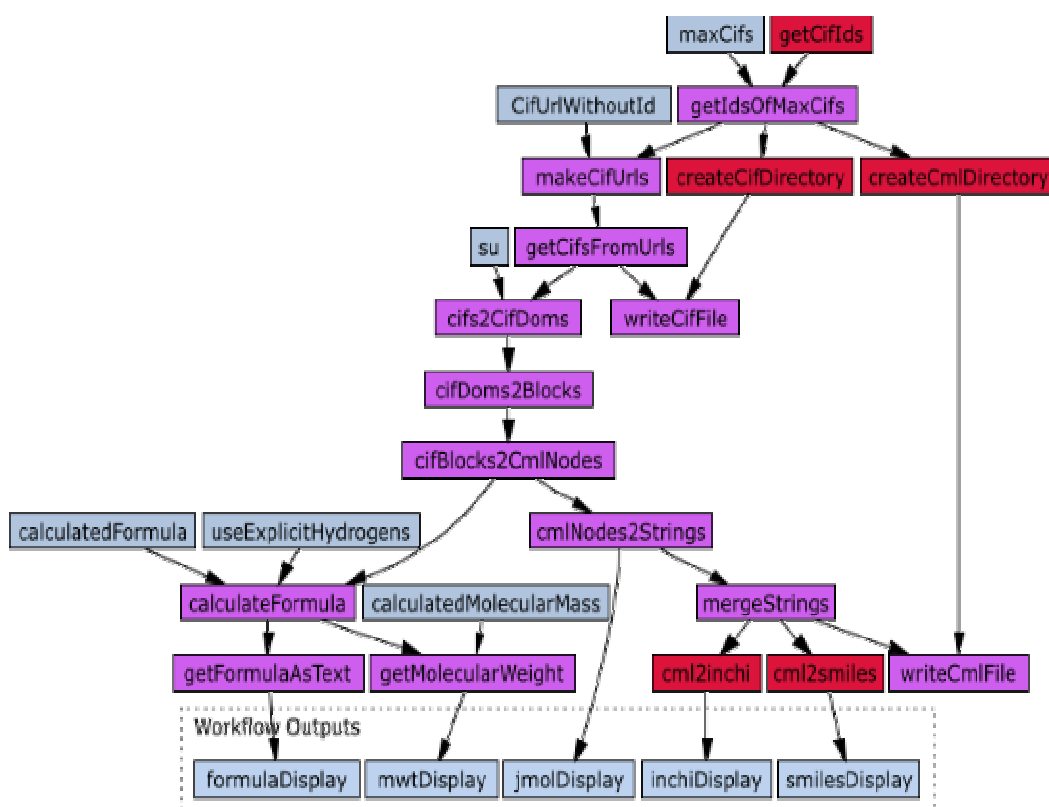


Figure 1. An automated workflow (built with the eScience/MyGrid Taverna tool) for adding chemical metadata to crystallographic data in repositories. It extracts data from the Int. Union of Crystallography's website and interprets the chemistry with over 97% precision.

2.7.2 Development of protocols. The greatest challenges are to make scientists aware of the need for deposition and to identify the perceived and real obstacles. Some current concerns are:

- if I publish my data it is available to other scientists who might make discoveries
- the data might be wrong and need amending
- I might want to patent or protect this at some stage
- it is too much work for no value

The benefits we can offer include:

- it is much easier to find data that existed in the lab and is now lost
- it makes preparation of theses and papers much easier
- it provides checks on quality at the time of data collection and not months/years later
- it may be mandated by funders or academic processes in the future
- it opens up new avenues of information-driven science.

2.8 Researching the current attitudes and aspirations of the chemical community will be an important part of the project and deliverables would include a re-usable protocol to help scientists (perhaps including training material).

- 2.9 In both Cambridge and Imperial College the respective libraries already have an institutional repository strategy in place.
- 2.10 Cambridge University Library, in conjunction with the University Computing Service, is managing and developing DSpace@Cambridge as its institutional repository.¹⁰ Implemented initially in collaboration with MIT Libraries (where the DSpace platform was first developed jointly with Hewlett-Packard), through a grant from the Cambridge-MIT Institute, DSpace@Cambridge has already acquired experience in handling a range of content types, ranging from research papers to datasets across a variety of academic disciplines. Along with the formulation of institutional policies and integration with other technical services, the DSpace@Cambridge team have also made a significant contribution to the Open Source development of the DSpace code and are collaborating with MIT Libraries on further development work in the areas of digital preservation and learning management systems.
- 2.11 Imperial College London Library is a member of the SHERPA-LEAP University of London SHERPA consortium and has an institutional repository running Eprints software on a shared server based at University College London. The repository scope currently covers eprints in the Physical Sciences. Having gained experience of repository related technical, policy and procedural issues through this project Imperial's strategy is to also establish its own separate repository on a College server using DSpace software and to extend the scope of the repository to include eprints, theses, research papers, and supporting datasets, for which the business case has already been prepared, across a number of scientific disciplines. It will also integrate the repository with other College information system, such as the Publications Database, RAE systems and the College Professional Web pages.

3 Partnerships

- 3.1 SPECTRA will build on work undertaken by the eBank project. We will work closely in collaboration with members of eBank, especially at UKOLN and Southampton. We envisage this partnership as including regular (at least three-monthly) joint meetings of SPECTRA and eBank representatives; due acknowledgement by both projects of the collaborative nature of their work in dissemination activities; a harmonisation statement to be issued jointly by the two projects detailing the nature of their partnership; and eBank representation on the SPECTRA project advisory board.
- 3.2 There are opportunities for complementary activity between SPECTRA and two other projects being submitted to the current JISC call and involving the SPECTRA partners. StORe proposes to undertake surveys of data-repository/output repository relationships in various disciplines, and the study of chemistry would be provided by Imperial. RAPIDO involves a study by UCL of the nature and types of learning objects being created in selected disciplines, including chemistry, and Cambridge would be contributing expertise gained from its experience with DSpace. Individually, both Cambridge and Imperial are also partners in other proposals relating to repositories but not specifically to chemistry. We will welcome the opportunity to collaborate with other relevant projects if JISC sees this as appropriate.

References

- ¹ http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf
- ² http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1,00.html [Annex 1]
- ³ Hey, T. & Trefethen, A.: 'The data deluge: an e-Science perspective'. In 'Grid Computing: Making the Global Infrastructure a Reality' ed. F.Berman et al. Wiley, 2003.
- ⁴ Murray-Rust, P., Rzepa, H.S., Tyrrell, S.M., and Zhang, Y.Y. (2004). *Representation and use of Chemistry in the Global Electronic Age. Org. Biomol. Chem.*, 2 (22), 3192 – 3203. DOI: 10.1039/b410732b
- ⁵ Day, N.E., Murray-Rust, P., Rzepa, H.S., Tyrrell, S.M., Zhang, Y.Y. (2005), *Enhancement of the Chemical Semantic Web through the use of InChI Identifiers Org. Biomol. Chem.*, in press.
- ⁶ <http://www.dspace.cam.ac.uk/handle/1810/724>
- ⁷ Rzepa, H.S., Murray-Rust, P., Williamson, M.J. and Willighagen, E.L., (2004), *Chemical Markup, XML and the Worldwide Web. Part 5. Applications of Chemical Metadata in RSS Aggregators, J. Chem. Inf. Comp. Sci.*, 44, 462-469. DOI: 10.1021/ci034244p
- ⁸ <http://www.iupac.org/projects/2000/2000-025-1-800.html>
- ⁹ <http://www.ukoln.ac.uk/projects/ebank-uk/>
- ¹⁰ <http://www.lib.cam.ac.uk/dspace/>