SPECTRa: The Deposition and Validation of Primary Chemistry Research Data in Digital Repositories

Jim Downing,[†] Peter Murray-Rust,[†] Alan P. Tonge,[*,†] Peter Morgan,[‡] Henry S. Rzepa,[§] Fiona Cotterill,[||] Nick Day,[†] and Matt J. Harvey[⊥]

Unilever Centre for Molecular Informatics, Department of Chemistry, Cambridge University, Lensfield Road, Cambridge CB2 1EW, U.K., Cambridge University Library, Cambridge University, West Road, Cambridge CB3 9DR, U.K., Department of Chemistry, Imperial College London, Exhibition Road, London SW7 2AY, U.K., and Imperial College Library and High Performance Computing Unit, ICT, Imperial College London, Exhibition Road, London SW7 2AZ, U.K.

Received December 23, 2007

The SPECTRa (Submission, Preservation and Exposure of Chemistry Teaching and Research Data) project has investigated the practices of chemists in archiving and disseminating primary chemical data from academic research laboratories. To redress the loss of the large amount of data never archived or disseminated, we have developed software for data publication into departmental and institutional Open Access digital repositories (DSpace). Data adhering to standard formats in selected disciplines (crystallography, NMR, computational chemistry) is transformed to XML (CML, Chemical Markup Language) which provides added validation. Context-specific chemical metadata and persistent Handle identifiers are added to enable long-term data reuse. It was found essential to provide an embargo mechanism, and policies for operating this and other processes are presented.

1. INTRODUCTION

Over a million syntheses of new chemical compounds (and many repeat preparations) are reported each year. It is mandatory to accompany each of these with supporting experimental information, including much from modern instrumentation such as spectroscopy (1H, 13C and other nuclei Nuclear Magnetic Resonance (NMR), Infrared (IR) and Ultraviolet and Visible (UV/vis), circular dichroism, high- and low-resolution mass spectroscopy). In many cases it is also necessary to carry out X-ray crystallographic analysis of solids and increasingly to compute the properties of proposed compounds through quantum mechanical (QM) calculations. The initial motivation for these analyses is to determine the structures of unknown products and more usually to confirm the identity of an assumed product. However although this may be the immediate end point for the chemist at the time—the compound is what we think it is—it is now increasingly required that the data should be made available for peer review and for transmission to posterity. Thus a Ph.D. thesis or dissertation, for example, will contain visual diagrams or peak lists of all the spectra of compounds synthesized in the work or computed coordinates for modeled systems. Many journals insist on these analyses and calculations being made available as supplemental/supporting data for the reviewers and for readers who wish to verify the work or to use as reference data. A typical example of the value of reported analytical data is in the synthesis of hexacyclinol,[1] where some authorities questioned the identity of a reported product and, later, by using medium-level QM calculations of NMR data suggested an alternative interpretation for the structure.[2]

However the culture of managing this data does not normally extend to formal curation and archival. In the study reported here many chemists confessed to having lost data that would have been useful if properly preserved. Traditionally it has been possible to archive some of the work through paper copies (e.g., in theses), but normally only diagrams or transcribed peak lists are saved. This process is very lossy, and the paper archive is anyway laborious to reinterpret, especially if derived quantities such as peak heights and integrals are needed. In general neither instrument manufacturers or chemical software houses provide software that addresses the problem of medium- and long-term data preservation.

Recently there has been a major emphasis throughout the world for academic institutions to save and display their primary and published research outputs. Initially this has concentrated on the paper or “full-text”, normally in PDF which is archived in an Institutional Repository (IR). There has been much high-level activity and debate on how this should be supported and the role of scholarly publishers in facilitating the depositing of such outputs.[3] It is now generally accepted in most institutions that all peer-reviewed publications and the institution’s theses (Ph.D., M.Sc.) should not only be formally preserved but also disseminated in electronic form through repositories. These repositories contain not only the theses but also the associated metadata describing formats, languages, rights, access, and other bibliographic concepts. In addition, and especially for Web search engines,

[*] Corresponding author e-mail: apt24@cam.ac.uk.
[†] Department of Chemistry, Cambridge University.
[‡] Cambridge University Library, Cambridge University.
[§] Department of Chemistry, Imperial College London.
[||] Imperial College Library, Imperial College London.
[⊥] High Performance Computing Unit, ICT, Imperial College London.

10.1021/ci7004737 CCC: $40.75 © 2008 American Chemical Society
Published on Web 07/29/2008

it is important to enable domain-specific metadata for discovery and classification.

Although these initiatives are only a few years old, almost all institutions are now actively pursuing them and many will mandate the deposition by all researchers in all disciplines. However authors are often reluctant to make the effort to provide metadata. Recently the UK Joint Information Systems Committee (JISC) has invested in a large program to research into the most effective ways of developing digital repositories and to pump-prime early adopters.[4] The work reported here took place under the SPECTRa project,[5] which had the main tasks of

- Formally surveying (by questionnaire and interviews) the need for and the practice of preserving chemical data
- Building a software system of sufficient power and flexibility to test the proposed strategy
- Making recommendations about suitable protocols and practices for chemistry departments and related repositories

Chemistry departments of the size involved in this project typically provide a technical service for both crystallography and NMR spectroscopy, and most have access to institutional high-performance computing (HPC) facilities which provide effective access to standard computational chemistry tools. Our survey found that in a typical department these services routinely produce 300+ solved small-molecule crystal structures per year and between 50,000−100,000 NMR spectra (of which an increasing proportion (currently *ca.* 20%) are multidimensional). An increasing proportion of synthetic chemists also carry out modeling or simulation studies. It has been reported that 80% of all crystallographic data are never published,[6] and we estimate that in organic chemistry 99% of all spectra (which are essential for the full analytical characterization and understanding of chemical structures) are lost. Even among ostensibly computer-literate molecular modelers, retention of e.g. computed molecular coordinates is highly variable. 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 typical scenario is that departing graduate students hand over media containing proprietary format binary data to supervisors, and this information rapidly decays due to lack of predictable data structure and indexing). Additionally, much of the Supporting Information currently submitted for peer-reviewed publication does not provide access to this primary data; images of the recorded spectra or molecular structures are presented but not the original data points or atomic coordinates. Most of these intrinsically high-quality objects decay with a short half-life and may be irretrievable within five years of their creation.

2. DATA REPOSITORIES

Open Archives Initiative (OAI) compliant institutional repositories are potentially an effective means of capturing, preserving, and disseminating this data in accordance with Open Access principles. By adding chemical metadata (e.g., the new IUPAC International Chemical Identifier,[7] InChI) we can enable high precision and recall from Web-based search engines (Google, MSN, etc.) which harvest our repositories.

University libraries, which in most cases provide the management of institutional repositories, have begun to explore ways in which they can acquire a better understanding of researchers' workflows and needs, in order to develop both the technical features and organizational policy for repositories and thus encourage their academic communities to make use of such facilities.

In both Cambridge and Imperial College, the respective libraries already had an institutional repository strategy in place. Cambridge University Library, in conjunction with the University Computing Service, had installed and developed DSpace@Cambridge as its institutional repository,[8a] acquiring experience in handling a range of content types ranging from research papers to data sets across a variety of academic disciplines, and the repository team had also made a significant contribution to the Open Source development of the DSpace code. At Imperial College, a project is under way to archive reprints deriving from faculty publications, also using DSpace.[8b]

While most repositories are currently designed and used for archiving and disseminating "full-text", there has been recent emphasis on the concurrent need to save data.[9] The project set out to develop a data archiving system and chose to use the existing repository technology rather than develop from scratch. The DSpace system[10] can manage

- Manual ingest of papers (as single, but not compound, documents)
- Association of metadata (especially Dublin Core) with an article
- A unique persistent identifier (handle) for each component
- Adherence to the OAI-PMH (Open Archive Initiative - Protocol for Metadata Harvesting[11]) protocol for harvesting metadata
- A GUI for browsing, navigation, and text-based searches of the text and metadata

In an early study we showed that it was possible to archive computational chemistry[12] and with an automatic script deposited over 250,000 calculations in the Cambridge DSpace.

3. METHODOLOGY

Initial Approach. We therefore created an initial design for the deposition of chemical data sets derived from departmental analytical and computational support services in such a way that would be a natural part of the technical workflow. The intention was that the digital output from the instrument or computer could be validated by machine and to which general and domain-specific metadata could be automatically added. We hoped that it would be possible to create a simple, generally applicable, workflow which could be implemented largely through a standard set of software components.

The project chose to focus on three distinct areas of chemistry research−synthetic organic chemistry, crystallography, and computational chemistry. At the outset, it was envisaged that detailed protocols for chemists' workflows would be developed. However, preliminary investigations revealed that these were not required to identify user requirements for deposition tools in the synthetic and computational areas as it was only the end points of the experimental or calculation process that produce the high-quality data that the project was interested in capturing for preservation−that is, data which chemists would consider useful as part of a regular report, a thesis or for submission to peer-reviewed publication.

Chemistry Disciplines. *Synthetic Chemistry.* Synthetic chemistry is the application of synthetic methods and reactions to create new chemical compounds, a process which depends heavily upon spectroscopic analysis (e.g., nuclear magnetic resonance, infrared, ultraviolet, and mass spectroscopies) to characterize these new chemical structures. Initial interviews were conducted at Imperial with selected synthetic chemistry research leaders. Although the potential value of repository-based preservation was appreciated (compared with traditional paper-based storage of spectra or the short half-life (ca. 5 yrs) of proprietary binary spectra data formats saved on CDs), these discussions indicated a clear reluctance to consider their use unless an embargo procedure for unpublished or commercially sensitive material was available. This became a central focus for subsequent data management policy in the project.

Crystallography. The role of crystallography in a chemistry department is primarily to provide a service to confirm chemical structure of newly synthesized compounds. The aggregation of these solved structures into a central repository or database, such as compiled by the Cambridge Crystallographic Data Centre,[13] adds significant value to the information. We conducted detailed workflow analysis with departmental crystallographers, to identify all aspects of the crystallographers' experimental procedure and interactions with their client chemists, and from which a detailed set of software requirements for a Sample Manager was identified,[14] with the functionality based on the online UK National Crystallography Service facility at Southampton.[15] However, the social situation facing departmental crystallographers at our institutions was found to be quite distinct from that faced in providing an online national service, and in the end the former indicated that the only part of the process that was of interest was the deposition of the final result−the refined crystallographic structure.

Computational Chemistry. Computational chemistry provides a theoretical study of molecular structure and properties that may not be readily amenable to experimental study. The extremely varied nature of computational chemistry methods employed within the subdiscipline (ranging from large-scale simulations on biological macromolecules using empirical force fields, to the study of isolated small molecules and transition states with *ab initio* quantum mechanics calculations) would have required significant resources−at a level not available to the project−in order to conduct adequate surveying of the long-term data preservation needs of the Computational Chemistry community. We therefore adopted the pragmatic approach of currently limiting efforts in this area to providing tools to capture the output from *ab initio* Gaussian calculations, a recognized standard methodology in this area.

As a further development arising from the project, a facility to capture the results of large-scale undergraduate computational chemistry calculations, submitted to a Condor-managed network of processors, was initiated at Imperial College in a teaching context.[16]

Serialized Data Formats. The ideal method for developing content repositories is through depositing XML documents based on one or more domain-specific namespaces. Chemical Markup Language CML[17] developed by Murray-Rust (Cambridge) and Rzepa (Imperial) is the *de facto* approach to representing chemical data and documents of this sort. Therefore it was appropriate to develop content and metadata based on this approach. Unfortunately much of the current scientific data is in legacy format, and we therefore chose a number of these formats which could reliably be transformed into CML and which represented the most common types of chemical experiment. We identified the following nonbinary formats which are accepted as data standards within the various chemistry disciplines chosen for study:

- Crystallography: CIF files[18]
- NMR: JCAMP-DX[19] and MDL molfiles[20]
- Computational Chemistry: Gaussian Archive files[21]

Crystallography. Crystallographic Information File (CIF) is a standard machine-readable text file format for authoring, archiving, and exchanging crystallographic information developed by the International Union of Crystallography (IUCr). It contains the atomic coordinates of the refined structure as well as author metadata, experimental conditions, crystal data, and the atomic coordinates of the refined structure. It is produced and supported by all diffractometer manufacturers.

NMR Spectroscopy. NMR spectroscopy is the common spectroscopic method used across all chemistry domains. The IUPAC-supported JCAMP-DX file format for spectroscopic data exchange[19] was identified as being the appropriate format standard for file input to our repositories as the use of binary proprietary format from spectrometer manufacturers (e.g., Bruker, JEOL, or Varian) would give data that would become unreadable over a period of time because of changes to manufacturers' software. Our initial work in this area has been limited to the deposition of 1-dimensional NMR spectra, based on the v5.01 standard,[19b] as the format proposal[19c] for multidimensional data has not yet been published in final form. An added complication is the existence of 'flavors' of the standard, whereby files produced by the different manufactures spectrometers contain extensive proprietary-based annotations.

As these JCAMP-DX spectra files do not contain structural information−an essential component of any (future) searching strategy−the additional deposition of a chemical structure file was also required. The MDL molfile, a proprietary format which has become established as a *de facto* standard in chemical information,[20] was chosen.

Computational Chemistry. The Gaussian archive entry from the output of a calculation[21] contains the molecule specification or optimized geometry (in internal coordinates) and all of the calculation's essential results. It also records standard metadata (user, date, and program version) used for the calculation.

Data Validation. In addition to simply depositing the required chemistry data files, it was recognized that some process of automatic data validation would be required as part of the Quality Assurance process. In archiving scientific data there is a requirement for the preservation of the units of measurement. Chemical Markup Language[17] (CML) was

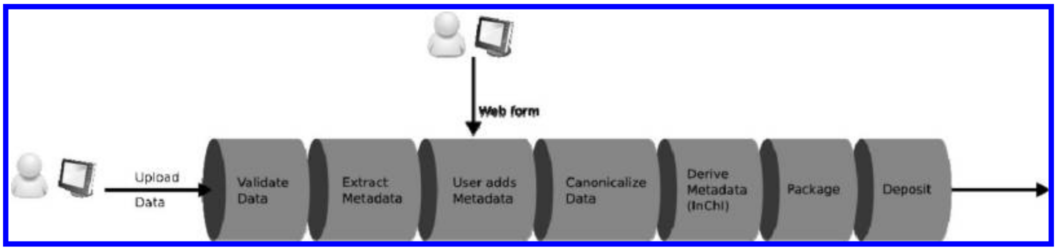


Figure 1. The data handling process.

designed to carry such units, and its self-identifying characteristics aids long-term preservation in a way that legacy formats often do not. It has given an opportunity to provide machine-based validation of marked-up chemistry data through the use of XSD schemas, which can set limits on data values and data ranges. Therefore all four file types identified above are converted to the appropriate CML subtype and data types validated programatically; any data type errors are caught and must be corrected offline by the user before deposition can be resumed. Some automatic data adjustment (e.g., for crystallographic disorder) is incorporated.

4. SOFTWARE ARCHITECTURE

The system was built on the basis of interoperable components for a subset of currently available open tools. For the three areas of chemistry examined, we used three existing legacy to CML converters, and for the repository a leading Institutional Repository software. We have designed extension points in the software so that other legacy applications and repository softwares can be included in the future. The submission and repository systems are packaged and openly available from SourceForge[22] and are described in the section below. In Cambridge we have also implemented a search system providing metadata and text search over chemical names, which is also described below.

Submission Web Application. The main software component delivered was a Web application to make submission of data along with necessary provenance metadata as straightforward as possible. Regardless of the types of data handled, the process follows the same stages, as shown in Figure 1.

The application supporting this process is a Java Servlet Web Application,[23] based on the Tapestry framework[24] (page templating, mediating between HTTP protocol and application components) and the Spring[25] component container (component lifecycle management, configuration, and dependency injection). It can be deployed in any standard Servlet engine (e.g., Tomcat,[26] Jetty[27]). The application is designed around reusable components orchestrated by the component container and a particular application. A configuration management system is provided to allow power users to repurpose the application from an XML file without needing knowledge of all the various Application Programmer Interfaces (APIs) used by the application itself.

Data File Handling. Support for different formats of data file is abstracted to make it easy to extend the application to different types of data. As an example of the reusability of the file handling components, a command line tool (also available from the SourceForge Web site[22]) is provided that can validate, extract metadata from, and convert data files to CML. This demonstrates the potential for the components to be repurposed as a batch processing tool to automate bulk deposits of existing content. The data file handling components have made use of a number of existing and tested Open Source libraries as follows:

CIF. SPECTRa uses CIF2CML to validate CIF files and convert them to CML. CIF2CML is a Java library based on CIFXML, a component of Jumbo.[28] It uses heuristics to extract chemical bonding including bond orders and conventional wedge/hatch stereochemistry from the crystallographic coordinates. This is not always possible for disordered structures, and in addition it is often impossible to calculate the charges on moieties if not given by the authors. Inorganic structures and some organometallics (especially infinite structures) cannot be reliably converted to InChIs. Metadata is extracted using XPath[29] from the intermediate CIFXML format.

Gaussian Archive Data. Gaussian2CML[30] is a Java library based on Jumbo which reads the archive sections from most versions of Gaussian output files. Normally these contain complete internal or Cartesian coordinate data and records of energies and orbital properties.

JCAMP-DX. Metadata from JCAMP files is extracted directly by SPECTRa code. Validation and conversion are performed using the JCAMP-DX library.[31] This JCAMP-DX library was written by the Creon laboratory and others, donated as Open Source to Sourceforge, and converts the spectral data into CMLSpect.[32]

MDL Mol. MDL Mol files are used in the system to provide a simple connection table and set of coordinates for data that would otherwise lack it (e.g., JCAMP-DX). The files are validated and converted to CML using MDLConverter and the Jumbo library.[28] Missing hydrogens are added according to valency rules (also by Jumbo). Any additional metadata present is ignored.

CML. CML files are the canonical format for SPECTRa and, as such, are not converted to any other format. They are validated by Jumbo, and the JNIInChI[33] and C-InChI[7b] libraries allow an InChI to be calculated for the chemical represented.

Packaging and Repository Deposit. In the default configuration, the submission tool uses Metadata Encoding and Transmission Standard[34] (METS) as a package manifest format and uses ZIP to form package archive files. These packages are then deposited into a DSpace repository using the Lightweight Network Interface[35a] (LNI). The functions of packaging and deposit are managed by a plugin system that allows for the development of components to support other packaging standards and deposit methods. As an example, a simple component that deposits packages in a local file system is provided, so that the application can be used without needing a DSpace installation.

During the SPECTRa project, the JISC initiated a project to specify a repository software agnostic Deposit API,[35b]

Table 1. Submission Application Customization

customization	method
Web user interface	edited using simple HTML templates and Cascading StyleSheets (CSS)
localization (L10N)	Support for non-English languages and locales outside the UK can be supported by providing additional translation files.
new data formats	Requires the development of a file handling plugin in Java.
new metadata fields	Requires the development of a User Interface template and handling code in Java.
new packaging format or repository interface	Requires the development of a plugin in Java.

which subsequently evolved into project SWORD.[35c] The latter has now completed (as of October 2007), delivering a consistent Web service API across the ePrints, DSpace, and Fedora repository platforms. There are plans to implement a SWORD interface for the SPECTRa submission application, which will allow it to deposit into many institutional repositories without further customization.

Customization. Many facets of the submission application can be customized (Table 1) without programming expertise, and more with programming, but without requiring knowledge of the underlying frameworks.

DSpace. The original system uses a standard DSpace software as the embargo repository, to provide compatibility with the Cambridge and Imperial College IRs. We have found that the DSpace system has few advantages when dealing with chemistry data. We are therefore recommending that users take a more lightweight approach to developing embargo repositories.

The installed DSpace repositories were registered with the Handle Sytem[36] to enable persistent identifier resolution in the form of either http://hdl.handle.net/10042/id or, using the Digital Object Identifier[37] handle resolver, http://dx.doi.org/10042/id where e.g. 10042 is the registered publisher identifier unique to the repository and ID is a unique reference code assigned by DSpace at the time of deposition.

Search System. Institutional Repository software rarely (if ever) supports chemistry search and visualization techniques out of the box. We implemented this locally as a separate system, demonstrating that offering data harvesting functionality enables domain specialists to develop their own methods for search and visualization. This also avoids bloat in centralized services. We implemented a simple search system (searching on chemical name and author metadata) with results displayed using Jmol[38] (Figure 5). Independently we have shown that both substructure search and crystal parameter search are easy to implement in systems up to 100,000 structures (CrystalEye[39]).

5. RESULTS AND DISCUSSION

Surveying. Extensive surveying of all research chemists (postgraduate students, postdoctoral workers and academic staff) was conducted at both Chemistry Departments at Imperial College London and the University of Cambridge. In order to evaluate their current use of computers and the Internet, and also to identify specific data needs, a questionnaire of 28 parts was devised. Although returns were voluntary, an overall response of 22% was achieved. The distribution of responses was found to be representative of the populations as a whole, giving us some confidence in the validity of the results, details of which are reported separately.[40]

The major findings to arise from the survey questionnaire were

- Much data are not stored electronically (e.g. laboratory books, paper copies of spectra)
- A complex mixture of data file formats (particularly proprietary binary formats) is used
- A significant ignorance of digital repositories among both faculty and students
- A requirement for restricted access to deposited experimental data

Analysis of the results showed that NMR spectroscopy was the common spectroscopic method used across the different chemistry disciplines (synthetic organic, organometallic, inorganic/materials), and it was therefore chosen as the example for deposition tool development in synthetic chemistry.

Legacy Data. We did not initially appreciate the scale of nonconformance and changing standards for legacy file formats and data types. Both crystallographic CIF and spectroscopic JCAMP-DX file formats are in a state of continuing development, and older files often do not reach the more recent standards. Depositing data using both the crystallographic and NMR spectroscopic tools proved problematic for this reason. For example, out of a set of over 200 legacy CIF files from the Chemistry Department at Cambridge (chosen from those structures already published in peer-reviewed journals) approximately 30% showed a variety of parsing errors which had to be corrected manually.

As NMR spectra are by default saved in proprietary Bruker format at both Imperial and Cambridge, it was necessary to explicitly convert these to JCAMP-DX using Bruker's Topspin software.[41] The resulting files are heavily annotated with Bruker comment fields, which are not used as primary metadata and may be removed by XSLT stylesheet processing of derived CMLSpect documents. We did not have the resource to upgrade the validation procedures to cope with the proposed modificationc of the JCAMP-DX format to accommodate 2D data sets, which currently comprise approximately 20% of the service NMR spectra recorded at our institutions.

Our experience with the deposition of legacy data indicates that an active resource will be required to manage such change.

Embargo Repository. Initially we had assumed that the chemist and the leader of the service would agree that there was a point ("the golden moment") at which the researcher best understands the experimental process, possesses a comprehensive package of information to describe it, and is motivated to submit it to a data management process. During the discussions it became clear that there is often great concern in the community about releasing data prematurely. Some of this is a natural consequence of competitive science—not wishing others to have access to results before they have been completely published and accepted. The natural tendency for managing digital data is therefore similar to paper books—put it in a cupboard and only allow members of the group to see it before publication. However this often leads to data loss—group leaders, graduate students move on, and papers are never fully written up (perhaps only as preliminary communications). The data are then forgotten and become irretrievable or uninterpretable.

Discussions with chemistry research leaders then led to the idea of a restricted embargo procedure as a necessary requirement for the deposition of unpublished or commercially sensitive material. This was not foreseen in the original design, so the initial workflow was largely refactored to include an "embargo repository" as an additional component. This repository could be used to ingest the raw data into a "closed archive" where there would be no chance of premature publication. At some later stage agreed by the service and the chemist, the data and metadata would then be transmitted to the full institutional repository. Therefore, when data are deposited using SPECTRa deposition tools, the user is actively required to indicate both the length of the embargo period (0−3 years) and the default status (review or release) to be adopted at the end of the requested period. This information held in metadata enables control of public access to the data held in the repository.

Although the requirements of an embargo repository are somewhat different from a full IR (e.g., there is no requirement to expose it to public search engines), we chose to use an installation of the DSpace repository because it was a cost-effective solution for the project. In retrospect it is likely that a simpler technology would be more appropriate as it may require customization for local needs. One example of the latter technology was the portal developed for the computational chemistry module, and which served as an interface between the job submission to the computing facility and DSpace itself. Until the departmental and institutional protocol for the embargo is developed we have not fitted an automatic mechanism for transferring data to the IR and will do this by manually operated scripts when required (e.g., an appropriately located publish button located within the embargo repository environment).

Although the concept of a data embargo is not new to us—for example, it has been used for many years for commercially valuable protein structures held in the Protein Data Bank[42]—the extension into metadata is an innovative development. This approach, which anticipates the use of a 'push' process to transfer an object to an Open Access repository, differs fundamentally from a 'pull' mechanism such as that developed by the JISC-funded EThOS project[43] to harvest e-theses released from embargo.

Metadata. Usable metadata is a crucial component of a searching facility for digital repositories in order to enable data reuse. As much metadata as possible is created automatically when the deposited files for each of the three chemistry areas are read by the appropriate validation processes. In the case of computational chemistry, the metadata information can be completely generated in an automatic manner, and no human intervention is required.[44] As some fields are defined as mandatory (embargo period, author names) these are additionally prompted for by an AddMetadata page in the deposition process if missing; deposition cannot proceed until these fields are filled manually.

Metadata schemas adopted by the project are based on the extended Dublin Core schema[45] published by eBank for the eCrystals project,[15] and limited local extensions have been adopted to distinguish between the originating chemist data owner ('creator') and the spectroscopist/crystallographer ('contributor'). Embargo information is also encoded, and an example is shown in Figure 2.

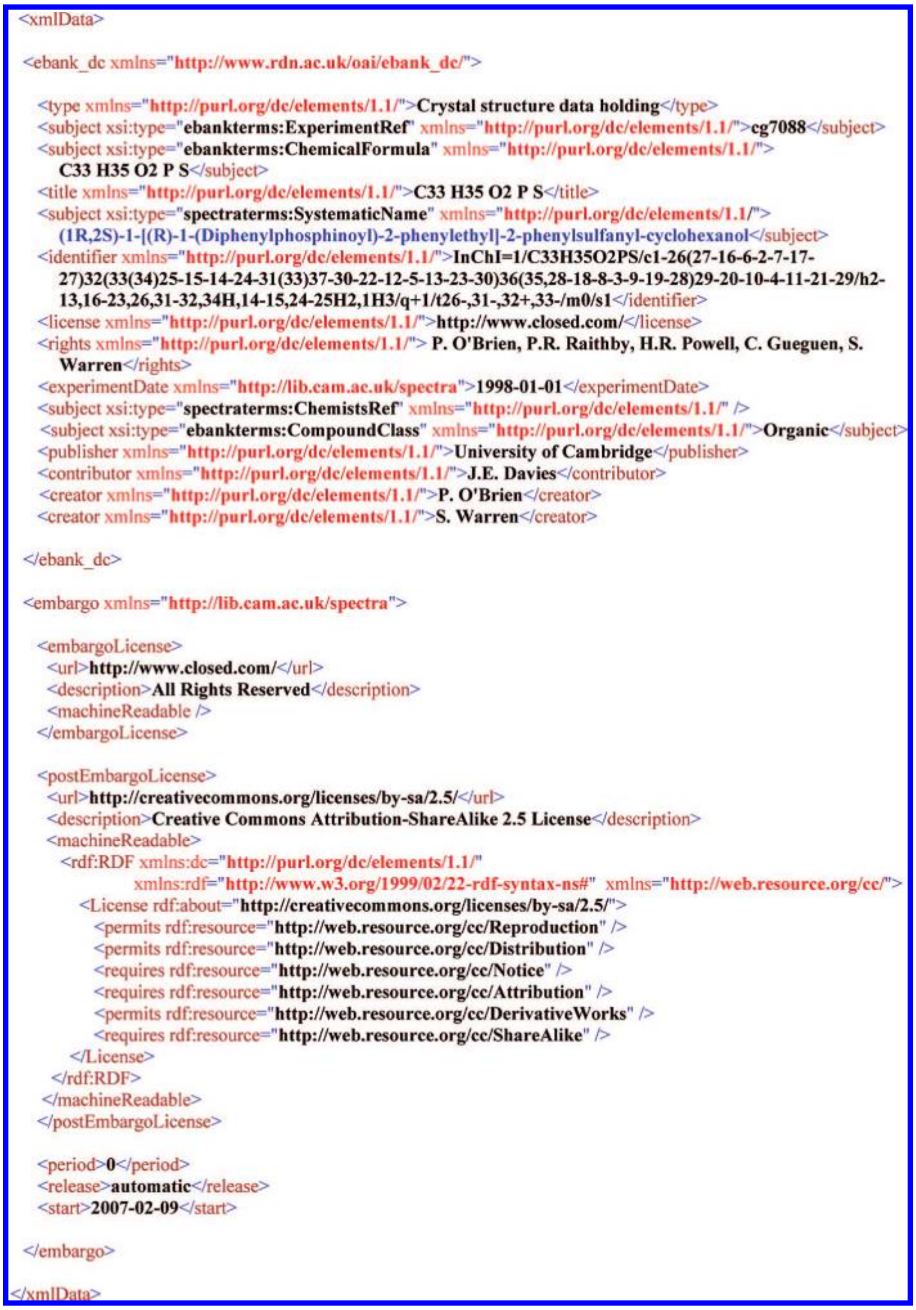
Data Preservation and Reuse. There are distinct types of data reuse. The project has provided tools which allow for the preservation aspects—*e.g.* a researcher can recover his/her data when writing up a thesis or report; other researchers can access items for individual inspection. However, additional tools would be required to add value to any large-scale data aggregates.

The value of semistructured documents in reporting chemistry information associated with publication is emerging, but there are as yet no standards or mandates for this.[46] The long-term preservation, quality assurance, and maintenance of chemical data in digital repositories is not a zero-cost option. Although a significant degree of automated data validation and metadata creation has been achieved through the use of dedicated software tools, there are limits to this (*e.g.* systematic name creation, author name identification), and some degree of explicit editorial support will have to be provided by future participating departmental and institutional repositories. For practical purposes we expect that serious work in preservation will be carried out at the institutional, rather than departmental, level. Nevertheless, we have made significant efforts to create both metadata and identifiers to enable long-term preservation, coupled with the provision of data in normalized open formats, including CML.

Preservation of chemistry data file formats is a difficult area. Anecdotal evidence (e.g., from departmental service managers) suggests that the expected lifetime of files in proprietary formats is around 5 years. For standards such as CIF the situation is considerably better, but since these files are generated from a range of different instruments and processes, the lifetime of older files may be shortened by variability in adherence to somewhat moving standards.

A developing program to evaluate and promote the long-term potential of chemistry-based digital repositories needs to be undertaken. As has been shown by the crystallographic community,[47] the potential added value of large-scale data sets in other areas of chemistry, such as NMR spectroscopy, is considerable. However, the scalability of institutional repository platforms in handling large data sets is still largely untested. This latter aspect of reuse was unfortunately beyond the resources of the current project but would be a necessary and important focus for future work in order to realize the full potential value of the deposited data.

Repository Deposition. There is much debate about whether repositories should be institution-based (e.g., DSpace@Cambridge[8]) or subject-based (including broad areas such as the biomedical field e.g. PubMed Central,[48] an online digital archive). There are no single answers, and much will depend on the source of research funding and institutional policy. We foresee a federated approach emerging. Within an institution the primary place for initial data capture will often be the department. The current model is to create a federation of repositories within an institution, coupled to the central institutional repository which will have special emphasis on aspects such as long-term preservation. Independently, by exposing their metadata, departments can create a federated subject repository with other same-subject



```
<xmlData>

 <ebank_dc xmlns="http://www.rdn.ac.uk/oai/ebank_dc/">

   <type xmlns="http://purl.org/dc/elements/1.1/">Crystal structure data holding</type>
   <subject xsi:type="ebankterms:ExperimentRef" xmlns="http://purl.org/dc/elements/1.1/">cg7088</subject>
   <subject xsi:type="ebankterms:ChemicalFormula" xmlns="http://purl.org/dc/elements/1.1/">
     C33 H35 O2 P S</subject>
   <title xmlns="http://purl.org/dc/elements/1.1/">C33 H35 O2 P S</title>
   <subject xsi:type="spectraterms:SystematicName" xmlns="http://purl.org/dc/elements/1.1/">
     (1R,2S)-1-[(R)-1-(Diphenylphosphinoyl)-2-phenylethyl]-2-phenylsulfanyl-cyclohexanol</subject>
   <identifier xmlns="http://purl.org/dc/elements/1.1/">InChI=1/C33H35O2PS/c1-26(27-16-6-2-7-17-
     27)32(33(34)25-15-14-24-31(33)37-30-22-12-5-13-23-30)36(35,28-18-8-3-9-19-28)29-20-10-4-11-21-29/h2-
     13,16-23,26,31-32,34H,14-15,24-25H2,1H3/q+1/t26-,31-,32+,33-/m0/s1</identifier>
   <license xmlns="http://purl.org/dc/elements/1.1/">http://www.closed.com/</license>
   <rights xmlns="http://purl.org/dc/elements/1.1/"> P. O'Brien, P.R. Raithby, H.R. Powell, C. Gueguen, S.
     Warren</rights>
   <experimentDate xmlns="http://lib.cam.ac.uk/spectra">1998-01-01</experimentDate>
   <subject xsi:type="spectraterms:ChemistsRef" xmlns="http://purl.org/dc/elements/1.1/" />
   <subject xsi:type="ebankterms:CompoundClass" xmlns="http://purl.org/dc/elements/1.1/">Organic</subject>
   <publisher xmlns="http://purl.org/dc/elements/1.1/">University of Cambridge</publisher>
   <contributor xmlns="http://purl.org/dc/elements/1.1/">J.E. Davies</contributor>
   <creator xmlns="http://purl.org/dc/elements/1.1/">P. O'Brien</creator>
   <creator xmlns="http://purl.org/dc/elements/1.1/">S. Warren</creator>

 </ebank_dc>

 <embargo xmlns="http://lib.cam.ac.uk/spectra">

   <embargoLicense>
    <url>http://www.closed.com/</url>
    <description>All Rights Reserved</description>
    <machineReadable />
   </embargoLicense>

   <postEmbargoLicense>
    <url>http://creativecommons.org/licenses/by-sa/2.5/</url>
    <description>Creative Commons Attribution-ShareAlike 2.5 License</description>
    <machineReadable>
      <rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
               xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"  xmlns="http://web.resource.org/cc/">
        <License rdf:about="http://creativecommons.org/licenses/by-sa/2.5/">
          <permits rdf:resource="http://web.resource.org/cc/Reproduction" />
          <permits rdf:resource="http://web.resource.org/cc/Distribution" />
          <requires rdf:resource="http://web.resource.org/cc/Notice" />
          <requires rdf:resource="http://web.resource.org/cc/Attribution" />
          <permits rdf:resource="http://web.resource.org/cc/DerivativeWorks" />
          <requires rdf:resource="http://web.resource.org/cc/ShareAlike" />
        </License>
      </rdf:RDF>
    </machineReadable>
   </postEmbargoLicense>

   <period>0</period>
   <release>automatic</release>
   <start>2007-02-09</start>

 </embargo>

</xmlData>
```

Figure 2. Marked-up Embargo and eBank crystallographic metadata (deposited with METS package).

departments and allow modern search engines to carry distributed subject searches transparently to the user.

The general architecture we have developed for repository deposition is shown in Figure 3, which highlights the separation of the short-term embargo processes required at departmental level from the longer-term preservation and Open Access requirements at institutional level. A more detailed description of the deposit/search tools is shown by the particular example of the crystallography toolset (Figure 4). Screenshots of the crystal structure search and display (Figure 5) and the computational chemistry deposition interfaces (Figure 6) are shown. Although basic repository search tools are distributed with DSpace, these have restricted display facilities, and additional text-based searching indexes have been built on the Web server which allows a richer range of search than that just based on metadata (InChI, systematic name) and authors.

Persistent Identifiers. The project had a requirement to provide a single identifier for content as it moves between systems (the deposition system, the embargo repository, the institutional repository) and also as a way to provide a link between the human readable and machine readable forms of the data packages generated. The lack of persistence of Internet Web pages is a common problem experienced by browsers, giving ‘Error 404’ messages which indicate that the requested Uniform Resource Location (URL) can no longer be found. We have chosen to use the Handle System,[36] and the URL assignment of a Handle identifier (e.g., hdl: 10042/to-568) has the format

e.g., http:// proxy_server/registrant_code/suffix
http://hdl.handle.net/10042/to-568

which allows the location of an entity (such as a Web page or, in our case, an object in a digital repository) to be changed, while maintaining the identifying name (a unique

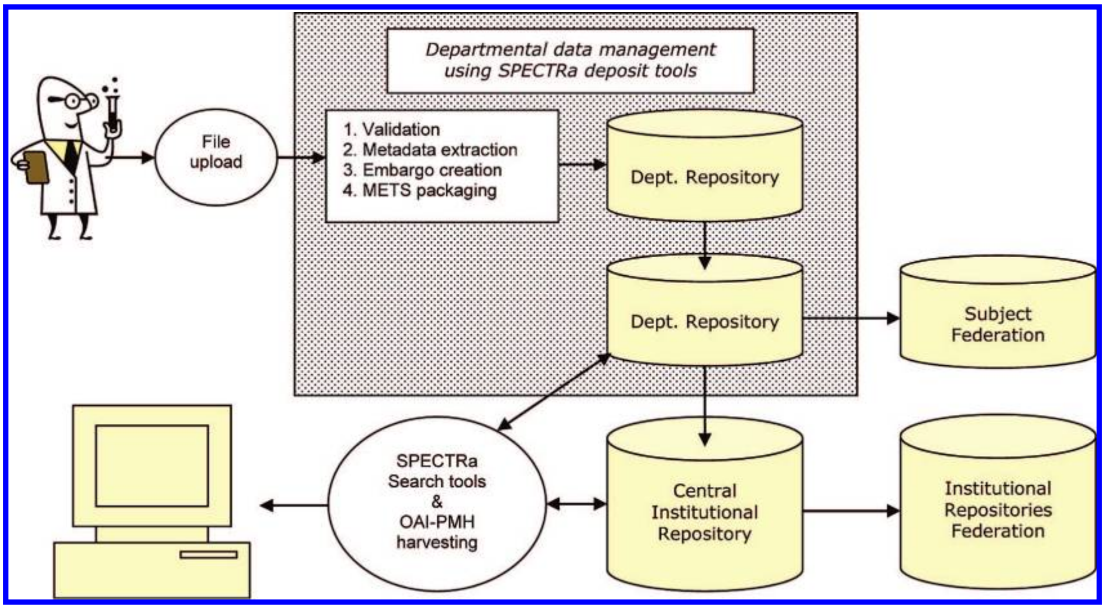

Departmental data management using SPECTRa deposit tools
File upload
1. Validation
2. Metadata extraction
3. Embargo creation
4. METS packaging
Dept. Repository
Dept. Repository
Subject Federation
SPECTRa Search tools & OAI-PMH harvesting
Central Institutional Repository
Institutional Repositories Federation

Figure 3. Repository deposition.

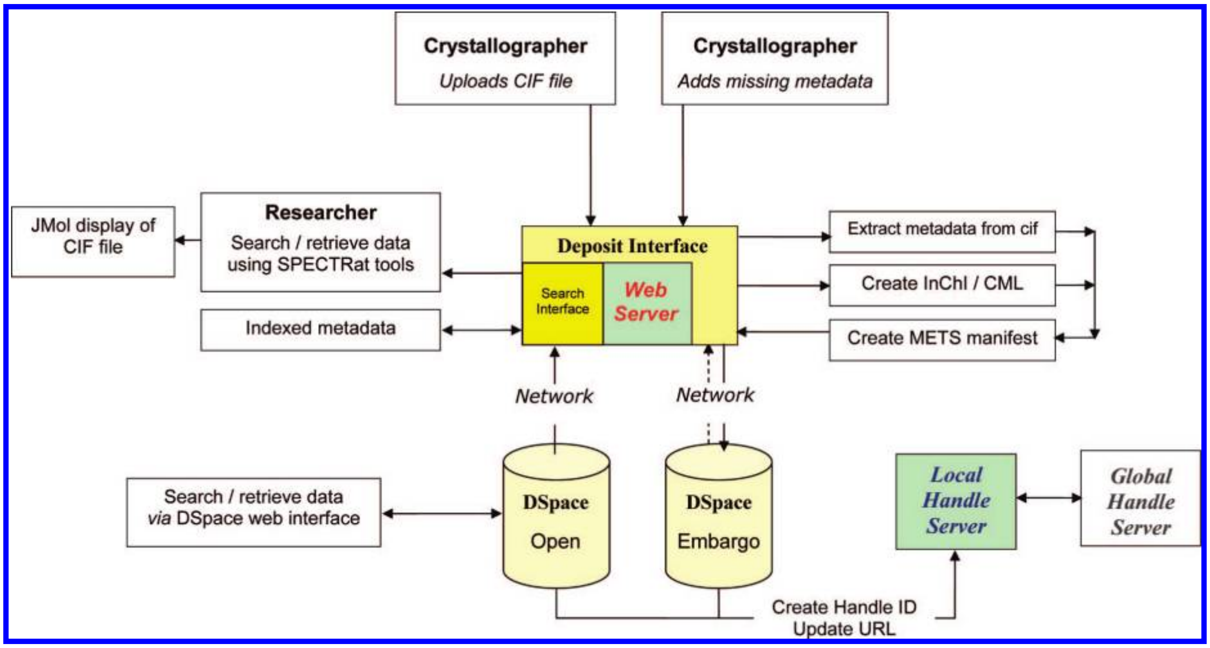


Figure 4. Architecture of crystallography deposition and search tools.

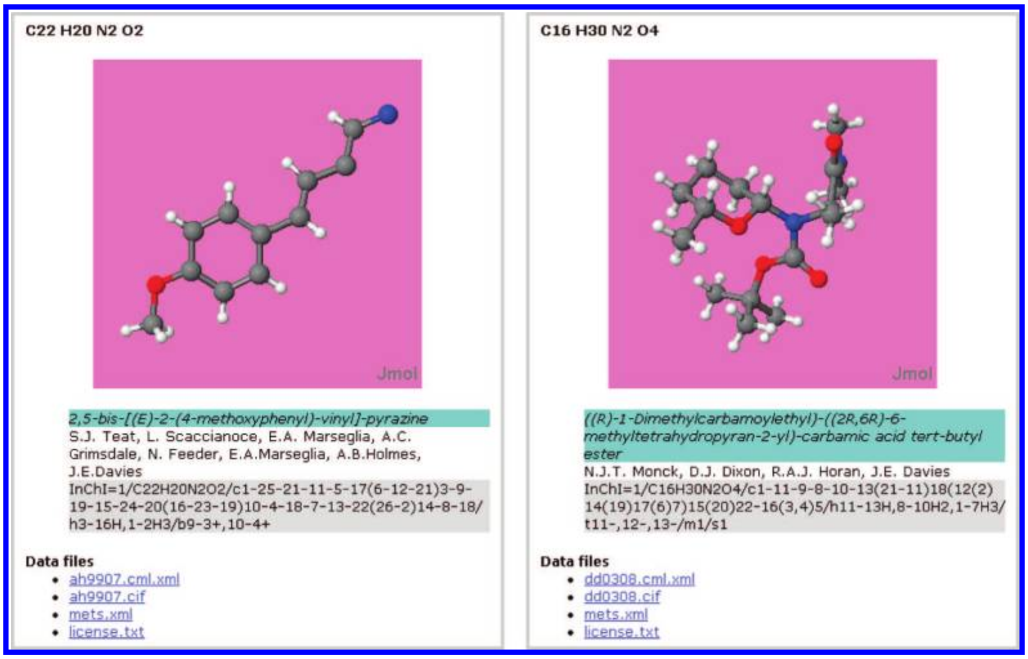
local identifier, indicated here by the suffix). The proxy server understands the Handle protocol and permits resolution of the location using a Web browser. All deposited packages, including those from the teaching application, therefore have a persistent Handle identifier attached to them. These identifiers are maintained in tables within DSpace.

Molecules and Other Linking Chemical Concepts. The emphasis in SPECTRa has been on archiving data related to single compounds ('molecules'). This reflects the current provision of analytical services in chemistry departments and the need for precise characterization. Since it is now possible to identify molecules algorithmically through the InChI identifier, a department can now *automatically* provide a unified view of all the measurements on a given compound. It would be relatively easy to extend the approach to other analytical or physical data, such as melting point or optical rotation, though these would need to be added manually by the chemists.

6. CONCLUSIONS AND RECOMMENDATIONS

Recognition of the problem of loss of primary research data is not new,[49] and our survey has shown that chemists still do not sufficiently appreciate the value of preservation of data in digital form.[40] Moreover, few have had experience of the processes involved. Many were unaware that institutions have digital repositories and that there is little active archival of primary publications or data. This may be, in part, because many chemical publishers forbid the archival of articles in public repositories.

Although we did not explicitly ask whether departments should provide a repository for data created by central services (particularly crystallography, spectroscopy and high-performance computing), there was an implicit agreement that, if this could be implemented painlessly, it would be a useful facility, for example, in gathering supplemental data for publication and for collating the material for inclusion



C22 H20 N2 O2

2,5-bis-[(E)-2-(4-methoxyphenyl)-vinyl]-pyrazine
S.J. Teat, L. Scaccianoce, E.A. Marseglia, A.C. Grimsdale, N. Feeder, E.A.Marseglia, A.B.Holmes, J.E.Davies
InChI=1/C22H20N2O2/c1-25-21-11-5-17(6-12-21)3-9-19-15-24-20(16-23-19)10-4-18-7-13-22(26-2)14-8-18/h3-16H,1-2H3/b9-3+,10-4+

Data files

- ah9907.cml.xml
- ah9907.cif
- mets.xml
- license.txt

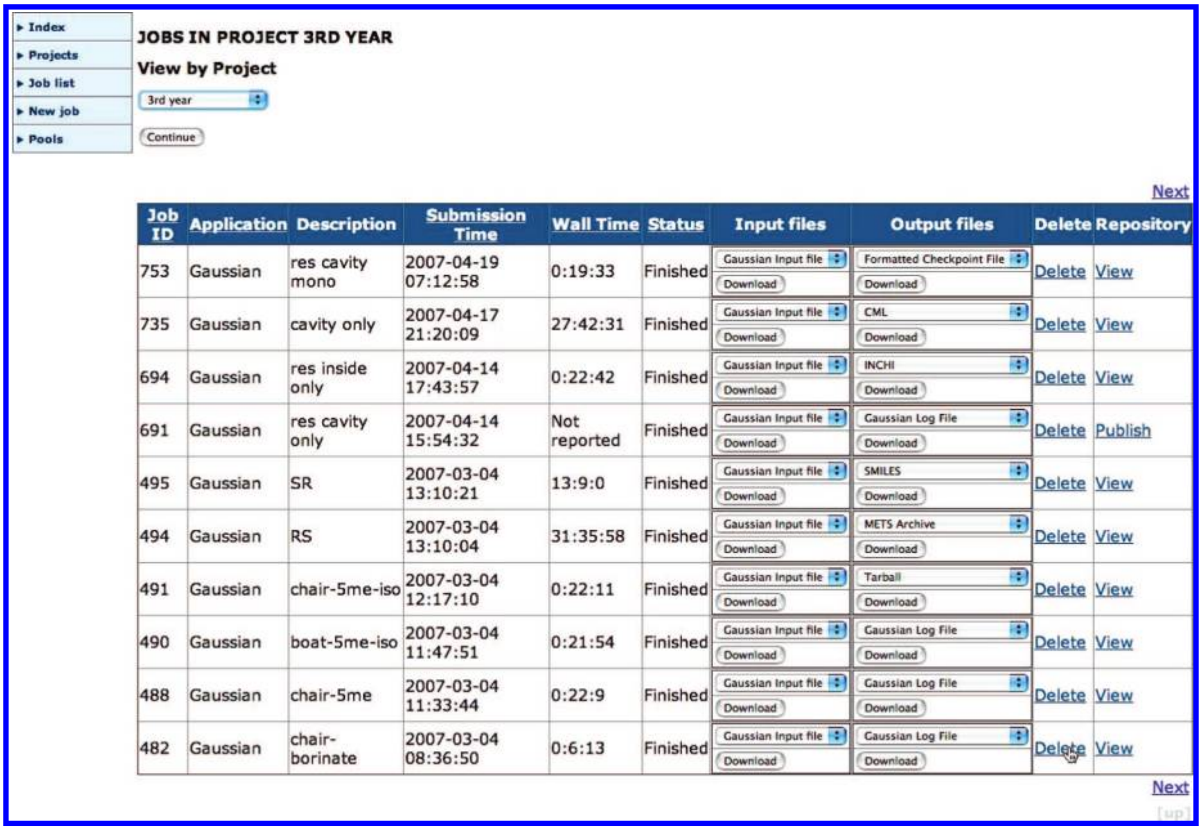
C16 H30 N2 O4

((R)-1-Dimethylcarbamoylethyl)-((2R,6R)-6-methyltetrahydropyran-2-yl)-carbamic acid tert-butyl ester
N.J.T. Monck, D.J. Dixon, R.A.J. Horan, J.E. Davies
InChI=1/C16H30N2O4/c1-11-9-8-10-13(21-11)18(12(2)14(19)17(6)7)15(20)22-16(3,4)5/h11-13H,8-10H2,1-7H3/t11-,12-,13-/m1/s1

Data files

- dd0308.cml.xml
- dd0308.cif
- mets.xml
- license.txt

Figure 5. Screenshot of the SPECTRa crystal structure search and display tool.



JOBS IN PROJECT 3RD YEAR

View by Project

3rd year

Job ID	Application	Description	Submission Time	Wall Time	Status	Input files	Output files	Delete	Repository
753	Gaussian	res cavity mono	2007-04-19 07:12:58	0:19:33	Finished	Gaussian Input file Download	Formatted Checkpoint File Download	Delete	View
735	Gaussian	cavity only	2007-04-17 21:20:09	27:42:31	Finished	Gaussian Input file Download	CML Download	Delete	View
694	Gaussian	res inside only	2007-04-14 17:43:57	0:22:42	Finished	Gaussian Input file Download	INCHI Download	Delete	View
691	Gaussian	res cavity only	2007-04-14 15:54:32	Not reported	Finished	Gaussian Input file Download	Gaussian Log File Download	Delete	Publish
495	Gaussian	SR	2007-03-04 13:10:21	13:9:0	Finished	Gaussian Input file Download	SMILES Download	Delete	View
494	Gaussian	RS	2007-03-04 13:10:04	31:35:58	Finished	Gaussian Input file Download	METS Archive Download	Delete	View
491	Gaussian	chair-5me-iso	2007-03-04 12:17:10	0:22:11	Finished	Gaussian Input file Download	Tarball Download	Delete	View
490	Gaussian	boat-5me-iso	2007-03-04 11:47:51	0:21:54	Finished	Gaussian Input file Download	Gaussian Log File Download	Delete	View
488	Gaussian	chair-5me	2007-03-04 11:33:44	0:22:9	Finished	Gaussian Input file Download	Gaussian Log File Download	Delete	View
482	Gaussian	chair-borinate	2007-03-04 08:36:50	0:6:13	Finished	Gaussian Input file Download	Gaussian Log File Download	Delete	View

Figure 6. Screenshot of the SPECTRa computational chemistry interface, showing a list of computations and their deposition status.

in theses. An additional utility which would encourage deposition into an IR would be the provision of a secure datestamping facility to establish the priority of a deposited article.

The project concentrated on the types of data which are very common in publications and theses. These are valuable not only for their role in validating the original science (usually a synthesis) but also as a resource for data-driven science. There have been thousands of publications on the correlation of crystal structures[50] and spectroscopy with molecular constitution, and by freeing large amounts of otherwise unobtainable data the departmental repositories can make an important contribution to chemistry. In the Internet era, it is routine for search engines to ingest the metadata and so provide a distributed global search facility for new chemistry.

Repositories will require an investment, and many institutions are addressing this. We recommend that capture of chemical data should be primarily the responsibility of the departments but that there should be a close liaison with the library and information service facilities to develop the best institutional infrastructure for curation and preservation. There are also specialist national facilities (e.g., the Digital Curation Centre in the U.K.[51]) which are able to give detailed advice. We also hope that national and international domain-specific organizations (learned societies, agencies) address how this infrastructure can be supported and maintained.

In this project we have created a set of Open Source components which we are now disseminating and which should be able to form the basis of a department's and institution's toolkit for preserving important chemical data. There is no "black-box" solution; every institution will have minor variations in procedures, and these will need to be built in locally. However the tools are built on standard Internet protocols and informed by the large amount of research into repositories, so it is likely that an institution will have enough local expertise to install and configure them.

IPR of Data. For the purposes of this project we chose to attach a Creative Commons license[52] to deposited data. This is consistent with Open Access principles and allows machine-readable copyright licenses. However, we do not wish to anticipate a general policy that other institutions and research funders should adopt. The lack of clear policies on IPR attached to data and its deposition in Institutional Repositories causes many problems. We recommend that explicit policies should be formulated and licenses should be given and have recommended to JISC that they prepare guidelines.

ACKNOWLEDGMENT

We thank the Joint Information Systems Committee (JISC) for funding and eBank for collaboration.

REFERENCES AND NOTES

(1) La Clair, J. J. Total Syntheses of Hexacyclinol, 5-epi-Hexacyclinol, Desoxohexacyclinol Unveil an Antimalarial Prodrug Motif. *Angew. Chem., Int. Ed.* **2006**, *45*, 2769–2773. DOI: 10.1002/anie.200504033.

(2) Rychnovsky, S. D. Predicting NMR Spectra by Computational Methods: Structure Revision of Hexacyclinol. *Org. Lett.* **2006**, *8*, 2895–2898. DOI: 10.1021/ol0611346.

(3) (a) Bailey, C. W., Jr. *Open Access Bibliography: Liberating Scholarly Literature with E-Prints and Open Access Journals*; Association of Research Libraries: Washington, D.C., 2005. (ISBN 1-59407-670-7). (b) 5th Workshop on Innovations in Scholarly Communication (OAI5), CERN, Geneva, Switzerland, 2007. (c) Suber, P. Open Access News. http://www.earlham.edu/∼peters/fos/fosblog.html (accessed April 4, 2008).

(4) JISC Digital Repositories Briefing Paper, 2005. http://www.jisc.ac.uk/uploaded_documents/HE_repositories_briefing_paper_2005.pdf (accessed April 4, 2008).

(5) (a) The SPECTRa Project was funded under the JISC 2005−2007 Digital Repositories Programme. http://www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories.aspx (accessed April 4, 2008). (b) Cotterill, F.; Downing, J.; Morgan, P.; Murray-Rust, P.; Rzepa, H. S.; Tonge, A. http://www.lib.cam.ac.uk/spectra/ (accessed April 4, 2008).

(6) (a) Davies, J. E. Crystallography News, 2004, Issue 88, p 22. http://img.cryst.bbk.ac.uk/BCA/on-line/2004/cn88m04.pdf (accessed April 4, 2008). (b) McMahon, B. Report and commentary on BCA chemical crystallography group meeting, 2004. http://bioportal.weizmann.ac.il/iucr-top/lists/epc-l/msg00757.html (accessed April 4, 2008).

(7) (a) InChI: The IUPAC International Chemical Identifier. http://www.iupac.org/inchi/ (accessed April 4, 2008). (b) Stein, S. E.; Heller, S. R.; Tchekhovski, D. An Open Standard for Chemical Structure Representation - The IUPAC Identifier. In 2003 Nimes International Chemical Information Conference Proceedings;Infonortics: Tetbury, Gloucestershire, U.K., 2003; pp 131−143. (c) Heller, S. R.; Stein, S. E.; Tchekhovskoi, D. V. InChI: Open access/open source and the IUPAC International Chemical Identifier. Abstracts of Papers, 230th ACS National Meeting, Washington, DC, United States, Aug 28−Sept 1 2005, CINF-060. (d) Coles, S. J.; Day, N. E.; Murray-Rust, P.; Rzepa, H. S.; Zhang, Y. Enhancement of the Chemical Semantic Web through InChIfication. *Org. Biomol. Chem.* **2005**, *3*, 1832–1834. DOI: 10.1039/b502828k.

(8) (a) DSpace@Cambridge. http://www.dspace.cam.ac.uk/ (accessed April 4, 2008). (b) Imperial College Digital Repository Spiral. http://spiral.imperial.ac.uk/ (accessed April 4, 2008) and https://spectradspace.lib.imperial.ac.uk:8443/dspace/ (accessed April 4, 2008).

(9) (a) Hey, A. J. G.; Trefethen, A. E. The Data Deluge: An e-Science Perspective In *Grid Computing - Making the Global Infrastructure a Reality*; Berman, F., Fox, G. C., Hey, A. J. G., Eds.; Wiley and Sons: 2003; Chapter 36, pp 809−824. (b) Lyon, E. Dealing with Data: Roles, Rights, Responsibilities and Relationships. http://www.ukoln.ac.uk/ukoln/staff/e.j.lyon/reports/dealing_with_data_report-final.pdf (accessed April 4, 2008).

(10) Tansley, R.; Bass, M.; Stuve, D.; Branschofsky, M.; Chudnov, D.; McClellan, G.; Smith, M. The DSpace Institutional Repository System: Current Functionality. In Proceedings of the 2003 Conference on Digital Repositories; IEEE, 2003; pp 87−97. http://hdl.handle.net/1721.1/26705 (accessed April 4, 2008).

(11) (a) The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html (accessed April 4, 2008). (b) JISC Information Environment Architecture OAI FAQ. http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/faq/oai/ (accessed April 4, 2008).

(12) Murray-Rust, P.; Rzepa, H. S.; Stewart, J. J. P.; Zhang, Y. A. Global resource for Computational Chemistry. *J. Mol. Model.* **2005**, *11*, 532–41. DOI: 10.1007/s00894-005-0278-1.

(13) Allen, F. H.; Battle, G.; Robertson, S. The Cambridge Structural Database In *Comprehensive Medicinal Chemistry II*; Triggle, D. J., Taylor, J. B., Eds.; Elsevier: Oxford, U.K., 2006; Vol. 3, Chapter 3, pp 389−410.

(14) Tonge, A.; Downing, J. Crystallographic Service Sample Manager. http://www.lib.cam.ac.uk/spectra/FinalReport.html (accessed April 4, 2008).

(15) (a) Coles, S. J.; Frey, J. G.; Hursthouse, M. B.; Light, Mark E.; Milsted, A. J.; Carr, L. A.; DeRoure, D.; Gutteridge, C. J.; Mills, H. R.; Meacham, K. E.; Surridge, M.; Lyon, E.; Heery, R.; Duke, M.; Day, M. An E-Science Environment for Service Crystallography - from Submission to Dissemination. *J. Chem. Inf. Model.* **2006**, *46*, 1006–1016. (b) eCrystals Structure Report Archive. http://ecrystals.chem.soton.ac.uk/ (accessed April 4, 2008).

(16) Rzepa, H. S. Molecular Modelling Workshop for Undergraduates, 2006. http://www.ch.ic.ac.uk/wiki/index.php/Second_Year_Modelling_Workshop (accessed April 4, 2008).

(17) (a) Murray-Rust, P.; Rzepa, H. S.; Wright, M. Development of chemical markup language (CML) as a system for handling complex chemical content. *New J. Chem.* **2001**, *25*, 618–634. (b) Murray-Rust, P.; Rzepa, H. S. Chemical Markup, XML and the Worldwide Web. CML Schema. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 757–772, and references therein.

(18) Hall, S. R.; Allen, F. H.; Brown, I. D. The Crystallographic Information File (CIF): A New Standard Archive File for Crystallography. *Acta Crystallogr., Sect. A: Found. Crystallogr.* **1991**, *A47*, 655–685.

(19) (a) Davies, N.; Lampen, P. JCAMP-DX for NMR. *Appl. Spectrosc.* **1993**, *47*, 1093–1099. (b) Lampen, P.; Lambert, J.; Lancashire, R. J.; McDonald, R. S.; McIntyre, P.; Rutledge, D. N.; Frohlich, T.; Davies, A. N. An extension to the JCAMP-DX standard file format JCAMP-DX V5.01. *Pure Appl. Chem.* **1999**, *71*, 1549–1556. (c) The draft form of JCAMP-DX v6.0 has been published as a PDF document on the Internet: Lampen, P.; Lancashire, R. J.; McDonald, R. S.; McIntyre, P. S.; Rutledge, D. N.; Davies, A. N. A Generic JCAMP-DX Standard File Format JCAMP-DX V.6.00. http://www.jcamp-dx.org/drafts/JCAMP6_2b%20Draft.pdf (accessed April 4, 2008).

(20) Dalby, A.; Nourse, J. G.; Hounshell, W. D.; Gushurst, A. K.; Grier, D. L. Description of several chemical structure file formats used by computer programs developed at Molecular Design Limited (MDL). *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 244–255. MDL is now part of Symyx Technologies, and the connection table manual has been updated accordingly. http://www.mdl.com/downloads/public/ctfile/ctfile.jsp (accessed April 4, 2008)

(21) Gaussian 03 Online Manual; Gaussian Inc.: Wallingford, CT 06492 U.S.A. http://www.gaussian.com/g_ur/k_archive.htm (accessed April 4, 2008).

(22) The SPECTRa Project. http://sourceforge.net/projects/spectra-chem/ (accessed April 4, 2008).

(23) Hunter, J. Java Servlet Programming, 2nd ed.; O'Reilly, 2001. http://java.sun.com/products/servlet/ (accessed April 4, 2008).

(24) Ship, H. L. Tapestry in Action, Manning (Pub), 2004. http://tapestry.apache.org/ (accessed April 4, 2008).

(25) R. Johnson, R.; Hoeller, J.; Arendsen, A.; Risberg, T.; Sampaleanu, C. Professional Java Development with Spring Framework, Wrox, 2005. http://www.springframework.org/ (accessed April 4, 2008).

(26) (a) Brittain, J. Tomcat: The Definitive Guide, 2nd ed.; O'Reilly, 2007. The Apache Software Foundation: Apache Tomcat. http://tomcat.apache.org/ (accessed April 4, 2008).

(27) Jetty Web server. http://www.mortbay.org/ (accessed April 4, 2008).

(28) Zhang, Y.; Murray-Rust, P.; Dove, M. T.; Glen, R. C.; Rzepa, H. S.; Townsend, J. A.; Tyrell S.; Wakelin, J.; Willighagen, E. L. JUMBO-An XML infrastructure for eScience, Proceedings of the UK e-Science All Hands Meeting 2004. http://archive.niees.ac.uk/documents/AHM_Jumbo_2004.pdf (accessed April 4, 2008).

(29) (a) Tidwell, D. XSLT, O'Reilly, 2001. (b) XML Path Language (XPath). http://www.w3.org/TR/xpath (accessed April 4, 2008).

(30) Downing, J.; Murray-Rust, P.; Rzepa, H.; Townsend, J.; Wright, M.; Zaharevitz, D.; Zhang, Y. Gaussian2CML. http://sourceforge.net/projects/cml (accessed April 4, 2008).

(31) Steinbeck, C.; Kuhn, S.; Weber, T. JCAMP-DX Software Project. http://sourceforge.net/projects/jcamp-dx/ (accessed April 4, 2008).

(32) Kuhn, S.; Helmus, T.; Lancashire, R. J.; Murray-Rust, P.; Rzepa, H.; Steinbeck, C.; Willighagen, E. Chemical Markup, XML, and the World Wide Web. 7. CMLSpect, an XML Vocabulary for Spectral Data. *J. Chem. Inf. Model.* **2007**, *47*, 2015–2034. DOI: 10.1021/ci600531a.

(33) Adams, S. JNI InChI Wrapper, 2006. http://sourceforge.net/projects/jni-inchi (accessed April 4, 2008).

(34) Cantara, L. METS: The Metadata Encoding and Transmission Standard. *Cataloging Classification Q.* **2005**, *40*, 237–253.

(35) (a) LightweightNetworkInterface. http://wiki.dspace.org/index.php/LightweightNetworkInterface (accessed April 4, 2008). (b) Deposit API. http://www.ukoln.ac.uk/repositories/digirep/index/Deposit_API (accessed April 4, 2008). (c) SWORD: Simple Web-service Offering Repository Deposit. http://www.ukoln.ac.uk/repositories/digirep/index/SWORD (accessed April 4, 2008).

(36) Handle System. http://www.handle.net/ (accessed April 4, 2008).

(37) Paskin, N. Digital object identifiers for scientific data. *Data Sci. J.* **2005**, *4*, 12–20.

(38) Hanson, R. M.; Willlighaven, E.; Vervelle, N.; Driscoll, T.; Howard, M. Jmol: an open-source Java viewer for chemical structures in 3D. http://jmol.sourceforge.net/ (accessed April 4, 2008).

(39) Day, N.; Murray-Rust, P. CrystalEye. http://wwmm.ch.cam.ac.uk/crystaleye/ (accessed April 4, 2008).

(40) Cotterill, F.; Morgan, P.; Tonge, A. SPECTRa Questionnaire Report: The Use of Computers and the Internet in Chemistry Research, 2007. http://www.lib.cam.ac.uk/spectra/questionnaire.html (accessed April 4, 2008).

(41) Bruker Topspin currently allows two JCAMP-DX formats to be saved: v5 and v6. However, the v5 does not include the published v5.01 revisionb which adopts the concept of an explicit shift reference; this is only incorporated into the v6 option for saving 2D data sets. Moreover, JCAMP-DX v6.0 (as already noted[19c]) is not yet published in final form, and the Bruker output includes a considerable amount of additional commented proprietary information. http://www.bruker-biospin.com/topspin.html (accessed April 4, 2008).

(42) Research Collaboratory for Structural Bioinformatics (RCSB). The RCSB Protein Data Bank. http://www.rcsb.org/pdb/home/home.do (accessed April 4, 2008).

(43) EThOS Electronic Theses Online Service. http://www.ethos.ac.uk/ (accessed April 4, 2008).

(44) Such calculations and the derived metadata have been deposited in an institutional digital repository at Imperial College London: Braddock, D. C.; Rzepa, H. S. Structural Reassignment of Obtusallenes V, VI and VII by GIAO-Based Density Functional Prediction. *J. Nat. Prod.* **2008**, *71*, 728–730.

(45) Duke, M.; Heery, R. eBank UK Feasibility Report on Dataset Description and Schema to access the results of experiments in crystallography, 2004. http://www.ukoln.ac.uk/projects/ebank-uk/schemas/feasibility/ (accessed April 4, 2008).

(46) Murray-Rust, P.; Mitchell, J. B. O.; Rzepa, H. S. Chemistry in Bioinformatics. *BMC Bioinformatics* **2005**, *6*, 141. DOI: 10.1186/1471-2105-6-141.

(47) (a) Helliwell, J. R.; Strickland, P. R.; McMahon, B. The role of quality in providing seamless access to information and data in e-science; the experience gained in crystallography. *Information Services Use* **2006**, *26*, 45–55. (b) McMahon, B. Semantically rich metadata in crystallographic publishing. http://www.iucr.org/iucr-top/lists/epc-l/pdf00003.pdf (accessed April 4, 2008).

(48) Roberts, R. J. PubMed Central: The GenBank of the Published Literature. *Proc. Natl. Acad. Sci.* **2001**, *98*, 381–2.

(49) (a) Heller, S. R. Where Have All the Data Gone? *Anal. Chim. Acta* **1982**, *136*, 1. (b) Casher, O.; Chandramohan, G. K.; Hargreaves, M. J.; Leach, C.; Murray-Rust, P.; Rzepa, H. S.; Sayle, R.; Whitaker, B. J. Hyperactive Molecules and the World-Wide-Web Information System. *J. Chem. Soc., Perkin Trans. II* **1995**, 7–11.

(50) Structure Correlation. Burgi, H.-B., Dunitz, J., Eds.; VCH: Weinheim, 1994; Vols *1*and *2*.

(51) Digital Curation Centre, University of Edinburgh, Edinburgh, U.K. http://www.dcc.ac.uk/about/ (accessed April 4, 2008).

(52) Creative Commons. http://creativecommons.org/ (accessed April 4, 2008).

CI7004737