

SPECTRa-T

Submission, Preservation and Exposure of Chemistry Teaching and Research Data from Theses

A proposal submitted by
the University of Cambridge (lead partner) and Imperial College London
to the JISC Capital Programme (call 04/06)

A. Introduction

- 1.1 This project will meet objectives in activity area (I), "Tools and Innovation", of the JISC 04/06 call for projects in its Repositories and Preservation programme, and is proposed by members of the current JISC-funded SPECTRa project (www.lib.cam.ac.uk/spectra). It will develop text-mining tools that address the need to extract the wealth of experimental data currently untapped in scientific theses, focussing on chemistry research data in molecular and related subjects. It will build on expertise gained in the SPECTRa, R4L and eBank projects, and on experience acquired by the Chemistry departments at the University of Cambridge and Imperial College London in Open Data publishing. The Cambridge group, an acknowledged world leader in chemical text-mining through the EPSRC SciBorg project (www.cl.cam.ac.uk/~aac10/escience/sciborg.html), has already produced widely disseminated tools that are being used, for example, in the high-throughput extraction of chemical information from PubMed abstracts.
- 1.2 Metadata specifications in this area have been developed by eBank and eCrystals, but a major challenge is the population of this metadata on a high throughput basis. Specialised humans are scarce and it is difficult to get author compliance with metadata deposition. To help this we shall automatically extract domain-specific data which can be used to create W3C SKOS (Simple Knowledge Organisation System) labels. A document such as a thesis will yield several thousand such labels and taken as a whole can be used to help identify the higher-level metadata which could be used to classify the thesis. In collaboration with eBank, 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 in a DSpace repository environment will be available as Open Source code designed for use with Open Standards-interoperable repositories.
- 1.3 The project will last for twelve months (April 2007 - March 2008). This timetable will enable it to run consecutively with Project SPECTRa, which is scheduled to end in March 2007.
- 1.4 The proposed project builds on, but is not a direct continuation of the current SPECTRa project. It assumes that an institution will have a SPECTRa-compliant means of ingesting data objects and also control over (probably including a mandate for) the deposition of electronic theses. It will investigate the utility of co-deposition of theses and data and will develop automatic methods to link experimental data with Knowledge Objects mined from the text and automatically linked to eBank metadata systems. We intend this will involve collaboration with the EThOS project and liaison with the National Centre for Text Mining (NaCTeM) in Manchester.

Aims

- 1.5 The aims of SPECTRa-T are to:
 - facilitate routine and automatic extraction of Knowledge Objects in high volumes, transformation into metadata and their ingest into institutional repositories.
 - survey current practice in the deposition of chemistry theses.
 - investigate the needs of the academic chemistry research community with respect to how data associated with theses may best be managed.
 - demonstrate how these needs may best be co-ordinated with emerging institutional strategies for repositories handling data-rich objects.
 - investigate the automatic discovery of data and data-rich documents in institutional repositories
 - investigate the cultural issues in capturing and re-using scientific data.
 - explore interoperability issues involving preservation of data in repositories.
 - develop semantic querying of institutional repositories.

Objectives

- 1.6 SPECTRa-T will realise these aims through the following objectives:
- review the practices of deposition of theses in crystallography, computational and synthetic chemistry.
 - integrate existing chemical KOS (IUPAC Gold Book) specifications into the deposition
 - develop automated validated and indexing tools specific to crystallography, computational chemistry and synthetic spectra, and providing interfaces with Open Standards-compliant repository platforms.
 - automatically increase the chemical KOS resource through text-mining.
 - disseminate and promote project outcomes to encourage widespread adoption.

B. Project description

- 2.1 John Wilbanks, Executive Director, Science Commons has observed: "Numerous scientists have pointed out the irony that right at the historical moment when we have the technologies to permit worldwide availability and distributed process of scientific data, broadening collaboration and accelerating the pace and depth of discovery.....we are busy locking up that data and preventing the use of correspondingly advanced technologies on knowledge"¹ The 2004 OECD Declaration on Access to Research Data from Public Funding has also 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 in furtherance of Open Data principles and practices, relating specifically to **data in scientific theses**. We will conduct a survey to establish what proportion (anecdotally estimated at 50%) is never published. Theses are an ideal target for mandating and assuring the capture of data with associated metadata because:
- 2.2.1 Institutions are moving towards the mandated deposition of e-theses (unlike – at present – primary publications).
- 2.2.2 The process is controlled to require certain degrees of quality both in content and metadata.
- 2.2.3 The student must always be in complete command of the experimental conditions and data when writing the thesis. Indeed a large part of many theses is made up of experimental data as tables, graphs and numbers.
- 2.2.4 The data are usually in machine-understandable form before being downgraded into text in (say) MSWord.
- 2.2.5 The younger generation, increasingly trained in good data-handling practices, is likely to be responsive to protocols for making their work machine-searchable.
- 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. Nevertheless the data (synthetic, spectral, computational and even crystallographic) generated in chemistry and related departments are conventionally attached to theses (e.g. PhD). In practice, as we have found in the current SPECTRa project, much of the data making up the scientific experiment and record is never communicated to the scientific community in appropriate form (numbers are reduced to points on diagrams, tables are converted to graphs in pixel form). For example a thesis of 200 pages might produce 2-3 papers of perhaps 5 pages each. The rest of the information– the data – never make it to the publisher. 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 Chemical information is essential to many sciences outside chemistry, including materials, life and environmental sciences, 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 annually in peer-reviewed publications from the global academic sector. The bare essentials of the synthesis are published but the detailed experimental recipe (as found in the thesis) is often 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). There is now great interest in extracting unpublished chemical compound information from theses (one of our team, PM-R, has been involved in early discussions with JISC, SURF [NL] and the NL company SORD on automatic extraction of chemistry from theses).

- 2.5 SPECTRa's final report will confirm that researcher compliance with deposition of data is a major obstacle. We have noted that the social aspects (ownership, fear of premature publication, etc.) were probably more important than the technical ones (e.g. lack of software) and militated against rapid deposition or high-compliance. A major achievement of SPECTRa has been the creation and testing of a "closed access archive" where researchers could deposit material in escrow. SPECTRa has also noted that there is a "golden moment" when the researcher and data provider agree on the scientific interpretation and quality of the data and when this could, barring social concerns, be reposted. SPECTRa has therefore developed a software protocol which is being used to gather metrics.
- 2.6 Theses have few of the social constraints mentioned above. A student must comply with regulations, must provide all supporting information to examiners if required, must assemble the data to a given quality metric, and must comply with escrow requirements. The creation of a thesis, therefore, represents a prolonged "golden moment" when data are in a perfect form to be reposted and where compliance can be assured.
- 2.7 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.
- 2.8 Methods for depositing this eChemistry are now being solved at a technical level and are being disseminated to HE departments, but need to be integrated more effectively to achieve their potential: this is a key element in SPECTRa-T. These will be available to those responsible for theses (librarians and heads of (chemistry) departments) and are freely available to students. The major technical developments in this project will be:
- 2.8.1 Streamlining and porting of the current SPECTRa technology. Our aim will be to develop a portable, free/Open Source, desktop tool into which bench chemists can dump their data. The tool will extract and systematise the types of data and extract Knowledge Objects in SKOS format. The tool will act as a "thesis checker" or "unit tester" which will advise the student when the appropriate data has been deposited and is of sufficient quality. (The preferred format is Chemical Markup Language – XML for chemistry – but we support conversion of major legacy formats (JCAMP, MDL-Molfile, SMILES, etc.) during the transitional period.
- 2.8.2 In addition we shall capture the complete recipe of the synthetic chemical procedure using text-mining developed in the SciBorg project. A typical recipe is shown in Figure 1 where the OSCAR3 software has identified data and chemical objects. We have systematized these into a KOS based on CML and expressible in the SKOS format.

9-Benzoyl-1-methyl-4-methylene-9-azabicyclo[3.3.1]nonane 18c

Following the procedure described for the preparation of **18a**, **17b** (310 mg, 0.95 mmol) was treated with **trifluoroacetic acid** (236 mg, 2.07 mmol) in DCM (5 cm³) and the crude product was chromatographed on **silica** gel [hexane–AcOEt (20 : 1)] to give **18c** (215 mg, 89%) as a colourless oil (Found: M⁺, 255.1630. C₁₇H₂₁NO requires M, 255.1623); ν_{\max} (film)/cm⁻¹ 1648; δ_{H} (300 MHz; CDCl₃) 1.57–2.19 (7 H, m), 1.69 (3 H, s, 1-Me), 2.29–2.46 (2 H, m), 2.65–2.78 (1 H, m), 4.34–4.37 (1 H, unresolved m, 5-H), 4.47 (1 H, br s, alkenic), 4.73–4.76 (1 H, m, alkenic), 7.33–7.43 (3 H, m, ArH) and 7.47–7.51 (2 H, m, ArH); δ_{C} (75.5 MHz; CDCl₃) 19.7, 30.3, 31.1, 31.5, 37.0, 38.8, 55.2, 60.6, 109.2, 127.5, 128.3, 129.9, 138.5, 146.7 and 173.7.

Figure 1. A typical chemical procedure indexed in the OSCAR system. Each term ("crude", "colourless oil", "trifluoroacetic acid") has been lexically (lexeme frequency) and linguistically recognized (e.g. with part-of-speech tagger) and interpreted. These form a large body of SKOS objects in the thesis, typically several thousand instances.

- A student may enter up to 200 such recipes in a thesis. Ideally the chemical objects (molecular connection tables) should be deposited (as in the SPECTRa protocol). This allows the system to be chemically searchable and classifiable using automatic tools.
- Where the student does not enter the connection table the OSCAR system is able to interpret a large number of the compounds by name lookup (lexicon of ca. 10 million names), or lexical reconstruction into the molecular formula.

- The physical properties (e.g. melting point) and analyses are converted to CML and thence to SKOS nodes. From a corpus with human assignments of metadata we shall construct an automatic classifier.
 - It is becoming increasingly common for experimental theses to also report properties computed using various modeling procedures, most particularly relative energies, geometries, charges, and conformational properties. These are particularly amenable to automated conversion into SKOS nodes.
 - The SKOS and metadata will be converted into RDF and thence into CML-RSS (a chemically-aware variation on RSS we introduced three years ago). We have already proofed the deployment of CML-RSS servers and shown that they can be used for transfer of objects.
- 2.8.3 Development of protocols. The greatest challenges are to make scientists aware of the need for deposition and to identify the perceived and real obstacles. As SPECTRa has demonstrated, researching the current attitudes and aspirations of the chemical community is essential in providing an evidence-base for future action to open up new avenues of information-driven science and to provide checks on quality of data.
- 2.9 In both Cambridge and Imperial College the respective libraries already have an institutional repository strategy in place.
- 2.9.1 Cambridge University Library, in conjunction with the University Computing Service, is managing and developing DSpace@Cambridge as its institutional repository¹⁰. Implemented initially in collaboration with MIT Libraries, DSpace@Cambridge has already acquired experience in handling a range of content types across a variety of academic disciplines. It is currently the largest DSpace instance in the world in terms of the number of files it contains, largely because of the content submitted by the Cambridge chemistry community. 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 continue to collaborate with MIT Libraries on further development work in the areas of digital preservation and learning management systems.
- 2.9.2 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 Library is also leading the implementation of an independent institutional repository at Imperial College, which will supersede the repository on the shared server, and will initially store the full text of academic publications. Further developments of the repository will include a DSpace instance for e-theses from January 2007. There is also the potential to store a wide range of academic output in the future.

3 Partnerships

- 3.1 SPECTRa-T will build on work undertaken by SPECTRa with the eBank and R4L projects, and we will continue to work closely in collaboration with colleagues in those projects.
- 3.2 There are opportunities for complementary activity between SPECTRa-T and other projects being submitted to the current JISC call and involving the SPECTRa partners. Imperial College Library is the lead institution in a repository proposal, EThOSNet, also being submitted under JISC call 04/06. The aims of this project are to build a strong EThOS sponsorship network, and to move the prototype EThOS service to a live, sustainable service. The work being proposed in SPECTRa-T may deliver outcomes that could add value to the core EThOS service, and we would welcome the opportunity to explore possible synergies between the two projects. We have also established links with CalTech and MIT with a view to obtaining additional theses from both institutions for testing purposes. We will welcome the opportunity to collaborate with other relevant projects if JISC sees this as appropriate.
- 3.3 We have made informal contact with NaCTeM, specifically in the area of chemistry. They have the general expertise for managing information extraction from text and Cambridge's EPSRC SciBorg project has created the chemical-specific component (OSCAR and OPSIN wwwm.ch.cam.ac.uk/wikis/wwwm/index.php/Oscar3).
- 3.4 Several publishers have been closely associated with SPECTRa, and the Royal Society of Chemistry in particular is keen to explore institutional data repositories. We have discussed potential value of capturing thesis data to support the publication process and will be using SPECTRa-T to gain expertise for a future bid to JISC in this area.

4 Work Packages

4.1 Work Package 1: Project management

Purpose: To provide overall project management, including: co-ordination of partner activities, liaison with JISC, and liaison with other relevant projects; testing, evaluation and reporting; financial management; communication, including project website.

Duration: Months 1-12.

Lead site: Cambridge

Deliverables: Project plan. Progress reports and final report. Financial reports. Website.

Formative evaluation.

4.2 Work Package 2: Analysis of stakeholder needs

Tasks: To scope project. To select early adopters in the testbed chemistry communities. To organise surveys of users' data-handling workflows and needs. To analyse the results and create specification for tools to be developed.

Duration: Months 1-6.

Lead site: Cambridge + Imperial

Deliverables: Reports. Tools specifications.

4.3 Work Package 3: Development of validation and indexing tools in chemistry

Tasks: To port desktop chemistry tools (Bioclipse plug-ins) and test their functionality in the new institutional contexts, and modify if necessary to ensure suitability. To productise OSCAR for chemical theses. To develop RDF-based tools to turn extracted terms into SKOS.

Duration: Months 1-12

Lead site: Cambridge + Imperial

Deliverables: Tested software. Automated chemical metadata creation. Knowledge Organisation Systems for chemistry.

4.4 Work Package 4: Protocols for deposition of data

Tasks: To develop protocols that can be embedded in research workflows for chemical theses, including workflows interfacing to high-performance computing (HPC) resources. To implement these among the testbed communities.

Duration: Months 3-12

Lead site: Cambridge + Imperial

Deliverables: Agreed protocols. Thesis creation tools.

4.5 Work Package 5: Dissemination and advocacy

Tasks: To promote the project and seek feedback through website, workshops and other events. To publish findings.

Duration: Months 7-12

Lead site: Cambridge + Imperial

Deliverables: Events. Publications. Community awareness.

4.6 Work Package 6: Metrics and evaluation

Tasks: Creation of metrics. Formative evaluation processes throughout project. Summative evaluation at end of project.

Duration: Months 1-12

Lead site: Cambridge + Imperial + independent consultant

Deliverables: Understanding of the project's findings and outcomes. Recommendations.

5. Timetable

Months	Project Management	Software Development
1-2 (WP 1, 2)	Scoping of project. Project advocacy among testbed communities. Selection of early adopters. Creation of metrics	Port and validate desktop tools and test functionality
1-6 (WP 1, 2)	Analysis of stakeholder needs. Collect requirements from senior synthetic chemists. Tools specifications.	Deploy search tools for chemistry to selected early adopters
3-12 (WP 1, 3)	Tool development.	Add chemical metadata/RSS/RDF functionality to tools. Productise OSCAR.
3-12 (WP 1, 4)	Protocols for deposition of chemistry in repositories	Implement chemical deposition system at both sites. Develop interfaces with high-performance computing resources

7-12 (WP 1, 5)	Disseminate project findings through conferences, publications, website	Disseminate project findings through conferences, publications, website
8-10 (WP 1, 3-4)	Supervise testing of deployed tools	Deploy spectral acquisition and search tools
10-11 (WP 1, 6)	Collect metrics and summarise findings. Presentation and feedback	Prepare distribution kit and documentation. Final presentation
11-12 (WP 1, 5-6)	Create protocols re-usable by JISC and HE. Summative evaluation by consultant. Final report	Finalise code release and documentation

6. Evaluation

- 6.1 The project will be evaluated as follows:
- Formative evaluation will take place throughout, and will be the responsibility of the Project Manager, who will report regularly to the Steering Group and the project team.
 - Periodic summative peer group evaluation will be conducted at intervals through workshops organised by the project.
 - Software tools will be tested in the course of the project at both partner sites.
 - Summative evaluation will be conducted in the final phase of the project by an independent external consultant.

7. Risk assessment

Potential problem	Probability	Risk management
Failure to develop and follow project plan	Low	Active project management and re-evaluation of progress by project team
Staffing: recruitment difficulties	Medium	Effective, targetted advertising. Existing staff able to provide temporary support.
Staffing: resignations during project	Low	Other staff in partner institutions have necessary expertise to avoid complete cessation of work
Financial	Low	Project manager supervised by experienced senior staff. Partner institutions ultimately liable for losses.
Technical: failure to develop tools as expected	Low	Staff expertise and familiarity with repository platforms should ensure success. Open Source community available to provide support.
Technical: failure of repository	Very Low	Implementation of institution's repository disaster recovery plan.
Organisational: lack of co-operation from departmental staff being surveyed	Low	Advocacy by project manager and departmental peer-group members
Organisational: problems between partner institutions	Low	Key personnel already have close working relationships
Legal	Low	Expert advice available from legal departments at Cambridge, Imperial, and JISC

8. Value of outcomes to JISC community

- Research students in chemistry: Data archival at time of capture gives higher quality, less loss (no cut and paste), increased discoverability within lab and institution. May dramatically reduce thesis creation effort (can take weeks to prepare data section for a paper).
- Those examining and storing theses. Will allow automatic checking of the technical correctness of a thesis (i.e. that all the bits are present)
- Researchers in other sciences: Methodologies and protocols can be used as exemplars elsewhere. Better comprehension of chemical concepts.
- Undergraduates: Exemplar of best practice in subject.
- Metadata creators: Large SKOS to use for chemical metadata and for general experimentation
- Libraries: Demonstration of opportunities for, and practical implementation of, library support for academic research. Opportunities developed for linkage between data and publications through repository interactions.

- Universities: Improved understanding of how institutional repositories can contribute to institutional strategies.
- Open Source community: Further contributions to the development and adoption of OS technology.
- Commercial sector: Improved exposure of scientific data and research outcomes will aid pharmaceutical and biochemical research and contribute to British industrial and economic performance. Publishers (and thus their authors and readership in academia) will benefit from improved practices in managing relationship between research papers and scientific data.

9. Intellectual property rights

- IPR in all materials created by the project will be held by the creators, who will exercise their rights in accordance with Open Access principles.
- JISC will be entitled to non-exclusive licences permitting it to utilise, archive and disseminate the outcomes as it requires.
- All software developed in Work Packages will be made freely available as Open Source code.

10. Sustainability

10.1 In the chemistry field we have built a substantive group in information extraction which is funded by research council, DTI and 3 publishers (likely to increase further). These publishers are interested in developing this technology further to solve the technical problems of archiving of scientific data at publication. We have collaboration with bioinformatics groups who see Open methods in chemistry as important and are likely to continue to develop in this area.

10.2 In both universities the libraries and computing services are already committed to developing institutional repositories as permanent features of our institutional knowledge management infrastructures. The University of Cambridge has approved a five-year business plan that ensures the continued development of the DSpace@Cambridge repository as a central infrastructure service.

C. Budget

11.1 Project SPECTRa - fEC Budget

Directly Incurred Staff Post, Grade & % FTE	March 07	April 07– March 08	April 08– March 09	TOTAL £
Project Manager, Cambridge (Senior Research Associate, 80% FTE)	█	██████	█	██████
Software Developer, Cambridge (Computer Officer Grade 8, 100% FTE)	£█	██████	█	██████
Software Developer, Imperial College (Level C, Point 43, 15% FTE)	██████	██████	██████	██████
Total Directly Incurred Staff (A)	£0	£94,567	£0	£94,567
Non-Staff	March 07	April 07– March 08	April 08– March 09	TOTAL £
Travel and expenses	£0	£2,000	£0	£2,000
Hardware/software	£0	£1,000	£0	£1,000
Dissemination	£0	£1,500	£0	£1,500
Evaluation	£0	£500	£0	£500
Other (Consumables, Printing)	£0	£400	£0	£400
Total Directly Incurred Non-Staff (B)	£0	£5,400	£0	£5,400
Directly Incurred Total (A+B=C) (C)	£0	£99,967	£0	£99,967

Directly Allocated	March 07	April 07– March 08	April 08– March 09	TOTAL £
Staff				
1. Principal Investigator, Cambridge (Senior Under-Librarian, 10% FTE)	£0	██████	£0	██████
2. Associate PI, Cambridge (Reader, 10% FTE)	£0	██████	£0	██████
3. Technical Support, Cambridge (Computer Officer, Grade 9, 5% FTE)	£0	██████	£0	██████
4. Associate PI, Imperial College (Professor, 10% FTE)	£0	██████	£0	██████
STAFF TOTAL	£0	£22,562	£0	£22,562
Estates	£0	£6,309	£0	£6,309
Other	£0	£0	£0	£0
Directly Allocated Total (D)	£0	£28,871	£0	£28,871
Indirect Costs (E)	£0	£47,704	£0	£47,704
Total Project Cost (C+D+E)	£0	£176,542	£0	£176,542
Amount Requested from JISC	£0	£99,967	£0	£99,967
Institutional Contributions	£0	£76,575	£0	£76,575
Percentage Contributions over the life of the project		JISC 57%	Partners 43 %	Total 100%

11.2 Budget justification

- A Project Manager is required with appropriate management skills and expertise in the fields of chemistry and repositories, to be responsible for overall management, including project planning, co-ordination of activity at the partner sites and liaison with JISC, and analysis of stakeholder needs. We estimate that an 80% FTE post will meet these requirements.
- A 100% FTE software developer is required to with experience in the design, testing and implementation of Open Source tools that are compatible with Open Standards-compatible repositories.
- A part-time (15% FTE) software developer is required to develop protocols and workflows that will provide interfaces between chemical theses and high-performance computing resources.
- To support the activities of these project staff, the project will require a budget to meet travel and subsistence costs incurred in attending JISC meetings and other relevant conferences; organising events to promote and disseminate the work and outcomes of the project; commissioning an independent summative evaluation; and to cover incidental consumables and administrative costs.

11.3 Summary of qualitative and quantitative benefits to partner institutions

Both partner institutions expect that SPECTRa-T will:

- improve the quality of data in their researchers' chemistry theses.
- expose more data and thus improve the cost-effectiveness of experimental chemistry.
- encourage chemists, and through their example other researchers, to appreciate the value of depositing research materials in their institutional repositories, and thus increase the rates at which repositories are populated with content.
- develop local expertise in analysing research workflows, identifying improved protocols, and developing new tools.
- develop local expertise in working with Open Source and Open Standards products.
- strengthen the role of libraries as managers of institutional repositories by demonstrating their willingness and ability to develop services in response to researchers' needs.

- create opportunities for new synergies within each partner's research communities through the identification of shared problems and development of collaborative solutions.
- create opportunities for collaborations with external institutions.

D. Key personnel

12 University of Cambridge

12.1 Peter Morgan - Principal Investigator

Peter Morgan is Project Director for SPECTRa and was also Project Director for the DSpace@Cambridge collaboration between Cambridge University Library and MIT Libraries from 2003 to 2006. He is a member of the University Library's senior management team, where he combines the responsibilities of University Medical Librarian with a co-ordinating role for digital library activities. He is a member of the SHERPA Management Group, and has been variously involved as a speaker, organiser, and participant in Open Access and digital repository meetings both in the UK and internationally over the past five years. He is a member of the Research Information Network's Consultative Group for Librarianship & Information Science.

12.2 Dr Peter Murray-Rust - Associate Principal Investigator and Chemistry lead

Peter Murray-Rust is Reader in Molecular Informatics in the Unilever Centre at the Department of Chemistry. Murray-Rust and his group have worked for several years with leading publishers and related organisations to develop new ways of using primary manuscripts. They have developed Chemical Markup Language (CML)¹¹, now used by groups such as the European Patents Office. They have created lexicons for major chemical concepts; developed rulebases for context analysis of words, phrases and regular expressions; and created a Bayesian/ME tool for the detection of probable chemical compounds in free text.¹² Large collections of named entities have been used to mark up manuscripts and they have also developed tools for autonomous resolution of terms through extraction from key websites.⁴ They have demonstrated that the data extracted from these manuscripts can be successfully collated, aggregated and reused. This work has led to a complete toolkit for the support of CML (authoring, editing, transformation, rendering, etc.) The group has constructed XML/CML data repository and search technology and developed institutional repositories as a source and sink for chemical and pioneered RSS as a means of publishing non-textual information.⁷ PMR is a Co-investigator with Computer Science in a very recently funded 4-year EPSRC/eScience project on "Extracting the Science from Scientific Publications".

12.3 Alan Tonge - Project Manager

Alan Tonge is an experienced synthetic organic and information chemist, having worked for over 20yrs at Glaxo Research as both a medicinal chemist and a molecular modeller and more recently at Imperial College and MDL Information Systems in the development and management of chemical information tools using web-based and proprietary technologies. He has been SPECTRa Project Manager for the past 12 months.

12.4 Jim Downing - Software Developer

Jim Downing is currently the developer on the SPECTRa project, and a committer for the global, open source DSpace project. As such he has software engineering expertise in an open source environment, as well experience in the management of an open source community's communications and management. He has six years of experience in programming and business consultancy for informatics projects, from knowledge management systems to large scale metadata repositories, with expertise in a number of technical fields including RDBMSs, XML and RDF, all of which may be involved in work packages 3 and 4.

13 Imperial College London

13.1 Professor Henry S. Rzepa - Associate Principal Investigator and Chemistry lead

Henry Rzepa is Professor of Computational Chemistry in the Chemistry Department, Imperial College London. Research activities in the area of computational modelling of industrially relevant metal-catalysed reaction mechanisms (currently funded by the EPSRC) and the discovery of fundamental new forms of aromaticity have been coupled with developing and raising awareness of new methods for publishing such information in a data- and semantically-intact form. JISC-funded projects to implement these long terms objectives into electronic journals (Project director, CLIC elib project in collaboration with The Royal Society of Chemistry, 1995-1997) and into functionally rich environments (VChemlab, 1997-1998) have operated in parallel with fertile collaborations with the Murray-Rust group to develop and deploy XML-based languages such as

the CML family^{4,7} as part of the construction of a chemical semantic Web. Current projects include application of metadata/RDF-based chemical resource discovery in document collections following on from an earlier project termed ChemDig for harvesting chemical metadata from Web collections.¹³ HSR has actively requested archival of supporting electronic information as part of journal publications for more than thirteen years now, a process which has frequently led to new (albeit limited) infra-structures being developed by publishers for the purpose, such as the **enhanced web-object** introduced recently into American Chemical Society publications.

13.2 Matt Harvey - Software Developer

Matt Harvey joined the Imperial College HPC unit in mid-2005, and is responsible for the infrastructure supporting some 2000 nodes in the system. He previously worked on the software workflows for creating an integrated instrument control and informatics system in combinatorial materials research at University College London.

14 Governance

14.1 The project will be overseen by a Project Steering Group with the following membership:

Dr Paul Ayris, Director of Library Services, University College London (Chair)
Peter Fox, Librarian, Cambridge University Library
Clare Jenkins, Director of Library Services, Imperial College London
Dr Simon Coles, Manager, National Crystallographic Service, University of Southampton
Dr David James, ICT & Production Director, Royal Society of Chemistry Publishing
Dr Liz Lyon, Director, UKOLN
Dr Alma Swan, Director, Key Perspectives Ltd.
Dr Peter Murray-Rust, Unilever Centre for Molecular Informatics, Chemistry Department, University of Cambridge
Professor Henry Rzepa, Dept of Chemistry, Imperial College London
Peter Morgan, Project Director, Cambridge University Library
Dr Alan Tonge, Project Manager (secretary)

References

- <http://sciencecommons.org/>
- http://www.oecd.org/document/0,2340,en_2649_34487_25998799_1_1_1_1_00.html [Annex 1]
- 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.
- Murray-Rust, P., Rzepa, H.S., Tyrrell, S.M., Zhang, Y. (2004). *Representation and use of chemistry in the global electronic age*. **Org. Biomol. Chem.**, **2**: 3192 – 3203. DOI: [10.1039/b410732b](https://doi.org/10.1039/b410732b)
- Coles, S.J., Day, N.E, Murray-Rust, P, Rzepa, H.S, Zhang, Y. (2005). *Enhancement of the chemical semantic web through the use of InChI identifiers*. **Org. Biomol. Chem.**, **3**:1832–1834. DOI: [10.1039/b502828k](https://doi.org/10.1039/b502828k).
- <http://www.dspace.cam.ac.uk/handle/1810/724>
- Rzepa, H.S., Murray-Rust, P., Williamson, M.J., 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](https://doi.org/10.1021/ci034244p)
- <http://www.iupac.org/projects/2000/2000-025-1-800.html>
- <http://www.ukoln.ac.uk/projects/ebank-uk/>
- <http://www.dspace.cam.ac.uk/> and <https://spectradspace.lib.ic.ac.uk:8443/dspace/>
- Murray-Rust, P., Rzepa, H.S. (2003). *Chemical Markup, XML, and the World Wide Web. Part 4. CML schema*. **J. Chem. Inf. Comp. Sci.**, **43**: 757-772. DOI: [10.1021/ci0256541](https://doi.org/10.1021/ci0256541); Holliday, G.L., Murray-Rust, P., Rzepa, H.S. (2006). *Chemical Markup, XML and the Worldwide Web. Part 6. CMLReact; An XML vocabulary for chemical reactions*". **J. Chem. Inf. Mod.**, **46**: 145-157. DOI: [10.1021/ci0502698](https://doi.org/10.1021/ci0502698)
- Townsend, J.A., Adams, S. E., Waudby, C.A., De Souza, V.K., Goodman, J.M., Murray-Rust, P. (2004). *Chemical documents: machine understanding and automated information extraction*. **Org. Biomol. Chem.**, **2**: 3294-3300.
- Gkoutos, G.V., Leach, C., Rzepa, H.S. (2002). *ChemDig: New approaches to chemically significant indexing and searching of distributed Web collections*. **New J. Chem.**, **26**: 656-666.