ETD 2008: Spreading the Light
11th International Symposium on Electronic Theses and Dissertations
Robert Gordon University, Aberdeen, 4-7 June 2008

Extracting and re-using research data from chemistry e-theses: the SPECTRa-T Project

Peter Morgan[1]; Jim Downing[2], Peter Murray-Rust[2], Diana Stewart[2], Alan Tonge[2], Joe Townsend[2]; Matt Harvey[3]; Henry Rzepa[4]

[1]Cambridge University Library;
[2]Unilever Centre for Molecular Science Informatics, University of Cambridge;
[3]ICT, Imperial College London;
[4]Chemistry Department, Imperial College London

[1]pbm2@cam.ac.uk

ABSTRACT

Scientific e-theses are data-rich resources, but much of the information they contain is not readily accessible. For chemistry, the SPECTRa-T project has addressed this problem by developing data-mining techniques to extract experimental data, creating RDF (Resource Description Framework) triples for exposure to sophisticated Semantic Web searches.

We used OSCAR3, an Open Source chemistry text-mining tool, to parse and extract data from theses in PDF, and from theses in Office Open XML document format.

Theses in PDF suffered data corruption and a loss of formatting that prevented the identification of chemical objects. Theses in .docx yielded semantically rich SciXML that enabled the additional extraction of associated data. Chemical objects were placed in a data repository, and RDF triples deposited in a triplestore.

Data-mining from chemistry e-theses is both desirable and feasible; but the use of PDF, the *de facto* format standard for deposit in most repositories, prevents the optimal extraction of data for semantic querying. In order to facilitate this, we recommend that universities also require deposition of chemistry e-theses in an XML document format. Further work is required to clarify the complex IPR issues and ensure that they do not become an unwarranted barrier to data extraction and re-use.

Introduction

The establishment of Open Access digital repositories in higher education has been accompanied by a growing awareness of their potential role in managing institutional collections of e-theses. International initiatives to define and promote this role have concentrated on four main areas of activity: institutional procedures for the voluntary or mandated deposit of theses; retrospective digitisation of earlier printed theses; preservation techniques to ensure the long-term viability of digital files; and, building on these efforts, networked resource discovery services to make full-text e-theses readily available on open access.

Less attention has so far been devoted to developing effective text-mining techniques that can ensure optimal retrieval and re-use of the information contained within a thesis. Scientific theses are a particularly rich source of data. In chemistry, postgraduate researchers generate significant quantities of experimental data, most of which will be reported in their theses. A 200-page thesis might typically contain 50 novel chemical preparations. However, it may then give rise to two or three peer-reviewed papers reporting only a small representative selection of these preparations (and then often printing them in inappropriate formats dictated by current publishing practices that render the data effectively unusable). Most of the information and data in the thesis, estimated to be about 70-80% of the total, will never be published in any other form and instead remain locked in the thesis.

The SPECTRa-T project[1], a collaboration between the University of Cambridge and Imperial College London, was designed to address this problem. It brought together expertise in Open Data publishing within the chemistry departments of the two universities, and also explored further the relationship between scientific departmental data management activities and the role of the central institutional repository. Funded by the Joint Information Systems Committee, the project ran from April 2007 to March 2008.

Aims and Objectives

The principal aims of SPECTRa-T, adopting a proof-of-concept approach and focusing on chemistry research data in molecular and related subjects, were:

- to develop text-mining tools and processes for the automatic extraction of experimental research data (chemical objects and named chemical entities);
- to transform the extracted data into metadata and ingest them into data repositories and RDF triplestores, thus enabling RDF-based semantic querying of the contents;
- to review current document format practice in the deposition of chemistry e-theses and assess its effect on data extraction.

Methodology

File Formats

We identified five file formats that are currently routinely available for thesis deposition: LaTeX, Postscript, Adobe Portable Document Format (PDF), Microsoft Word (DOC), and

Microsoft Office Open XML (DOCX). We chose first to investigate theses in PDF because this has been widely adopted by repositories as the standard format for thesis deposition.

Our subsequent studies of PDF indicated that it was not amenable to the structural analysis necessary for identification of chemical objects such as molecules, spectra, and physical properties. To facilitate such analysis we required a marked-up document format and chose Office Open XML (.docx)[2].

Document Structure

Text-mining tools may be expected to extract relevant data more accurately from a chemistry thesis if its key structural components (such as Title Page, Table of Contents, Introduction, Discussion, Experimental sections, Abbreviations, and References) can be recognised. This would permit standard Dublin Core metadata elements to be identified, and would enable the tools to be applied to the data-rich experimental sections, ignoring others with no data or, like Abbreviations, containing character-strings that might be mistaken for chemical terms.

We found that some institutions, e.g. CalTech and MIT, mandate the structure of their theses thus making these theoretically more amenable to processing by applying institution-specific rules. However, the observed variability in structure between different institutions required a level of development for automated analytical processes that was beyond our resources.

Thesis Sourcing

We downloaded a total of approximately 100 chemistry theses available in PDF from repositories in the UK (St Andrews and Stirling) and from the USA (Caltech and MIT). Those acquired from MIT proved to be predominantly digitized texts created by Optical Character Recognition from print originals. There are acknowledged problems with OCR accuracy, resulting in misassigned characters, and we therefore removed these MIT theses from our test set.

To obtain marked-up texts we acquired approximately 20 theses held in Microsoft Word within the Chemistry Department at the University of Cambridge, and these were manually converted to Office Open XML using Word 2007.

Software tools

Our chosen text-mining tool was OSCAR3 (Open Source Chemistry Analysis Routines)[3], an application developed by the SciBorg project[4] at Cambridge using natural language processing and established chemistry-domain ontologies to identify chemical terms . Using a regular expression facility, it has the ability to recognise experimental sections in document text. It converts human-readable chemistry text into XML marked-up content which can be manipulated by computers. (Fig.1)

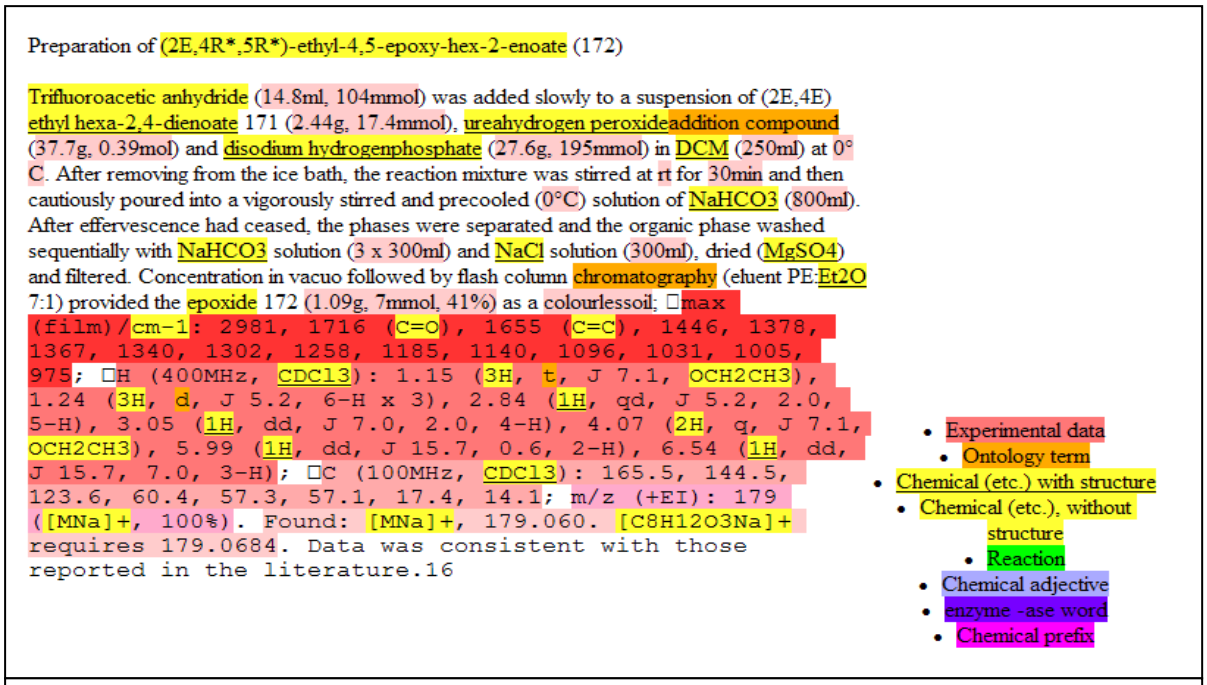

Fig.1: Highlighted experimental procedures created by OSCAR3

OSCAR3 is optimised to work with SciXML, typically derived from marked-up documents such as HTML, though it can also work with plain text. Chemical objects are more identifiable with a richer mark-up, and it was therefore desirable to provide SciXML for the document-processing. This can be obtained from both PDF and .docx documents: however, PDF delivers only a fairly basic SciXML output, wrapping the text in simple top-level elements, while .docx provides a much richer SciXML. The final output of the OSCAR3 process is SAFXML (Standoff Annotated Format XML), which can contain annotations referring to objects external to the XML document.

PDF processing

PDF (Page Description Format) is optimised for human, not machine, readability, and therefore describes the graphical appearance of page content. We found at least five problem areas affecting the suitability of PDF for text-mining:
- irregular word order;
- line breaks, with loss of continuous text and difficulty in identifying paragraphs;
- loss of subscript and superscript characters;
- non-printing characters;
- erroneous character assignment with OCR, indicating that born-digital PDF is preferable to OCR-scanned files

The loss of continuous text streams from the original document was a particular problem, for two reasons. Because systematic chemical names are often long, they may overlap onto the following line, but the line breaks imposed by PDF meant that OSCAR could not recognise the full chemical name separated into text-strings on two lines. Moreover, paragraph and

section endings were not identifiable, and preparative sections containing chemical objects could not be reliably recognised.

Before OSCAR could be used to mine chemical terms from a PDF document it was necessary to apply a number of text-processing tools in order to convert the PDF document automatically into usable SciXML. This process stripped out unwanted information such as font changes and images, and removed linefeed characters in order to reinstate continuous text streams. It utilised UTF-8 Unicode to preserve Greek characters (important as symbols in chemistry documents) that would be lost in simple ASCII text. Non-printing null characters and disconnected text derived from broken chemical structures and tables were also removed.

The resulting SciXML document was then analysed by OSCAR to retrieve named chemical entities (NCE's - commonly used names and expressions) that identify instances of essential chemical concepts, and XSL stylesheets transformed the results into RDF (Resource Description Framework) output. The basic unit of RDF is the triple - a statement containing a subject (or resource), predicate (property), and object (value) that can be searched using the RDF query language SPARQL. The triples resulting from this process were deposited in an RDF triplestore.

The workflow for processing PDF theses is outlined in Fig.2.

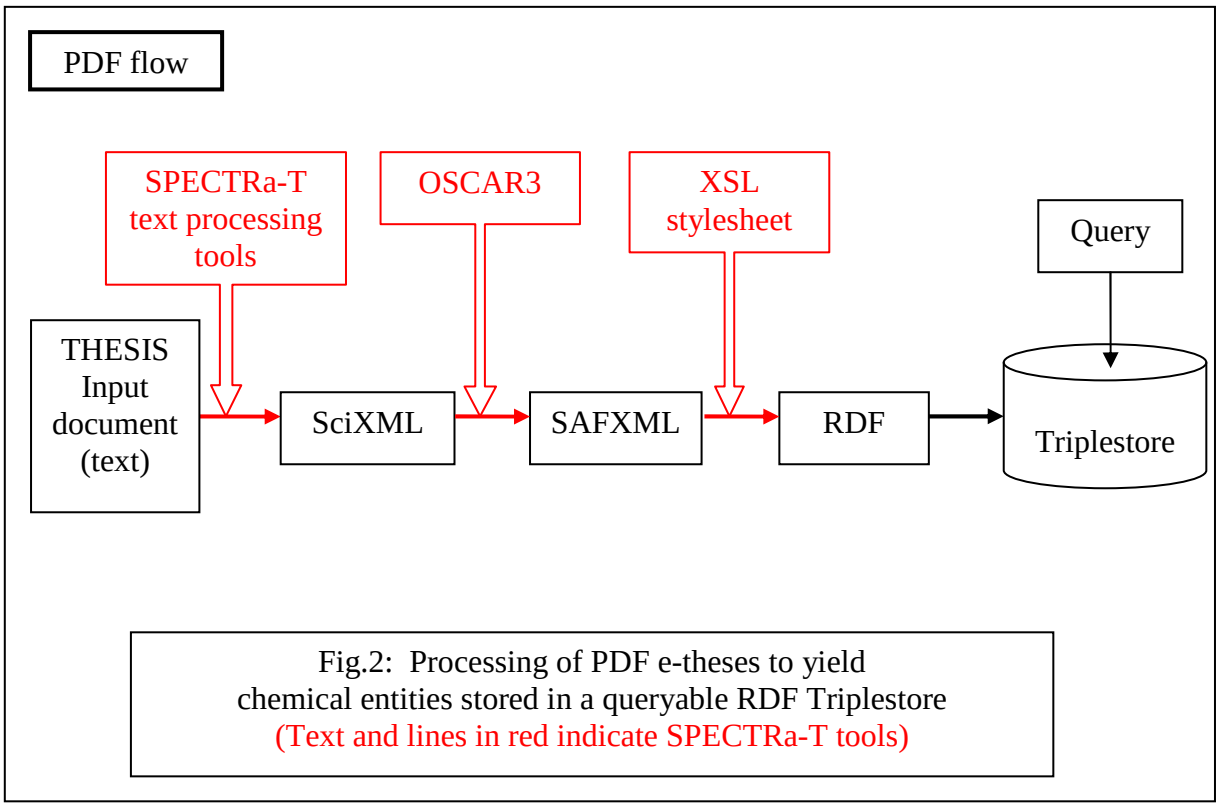

Fig.2: Processing of PDF e-theses to yield chemical entities stored in a queryable RDF Triplestore
(Text and lines in red indicate SPECTRa-T tools)

DOCX processing

As previously noted, the internal structure of PDF documents meant that the limited SciXML obtained from them prevented OSCAR from identifying chemical objects. The absence of these data would seriously have compromised our objective of creating a semantically-

queryable data repository. It was therefore important to obtain documents in XML that would offer greater scope for OSCAR analysis.

Theses in Word were converted manually to Microsoft Office Open XML using Word 2007 and then transformed into SciXML. Because .docx preserves paragraph structure (unlike PDF), it permits the identification of preparative sections. OSCAR is able to find these sections by locating the keyword "Experimental" in the SciXML document. Chemical objects can then be extracted and the resulting data XML files converted into CML (Chemical Markup Language)[5]. Each new preparation and the associated spectral assignment data is given a unique URI directory filename (indicating a webserver with an associated filestore) and placed in a data repository.

By adding the URI to named chemical entities in the SAFXML, using the name as a reference, they can then be updated with the location of these files, thus enabling semantic querying to link the derived RDF in the triplestore with the extracted CML files of chemical objects in the data repository.

The workflow for processing .docx theses is outlined in Fig.3.

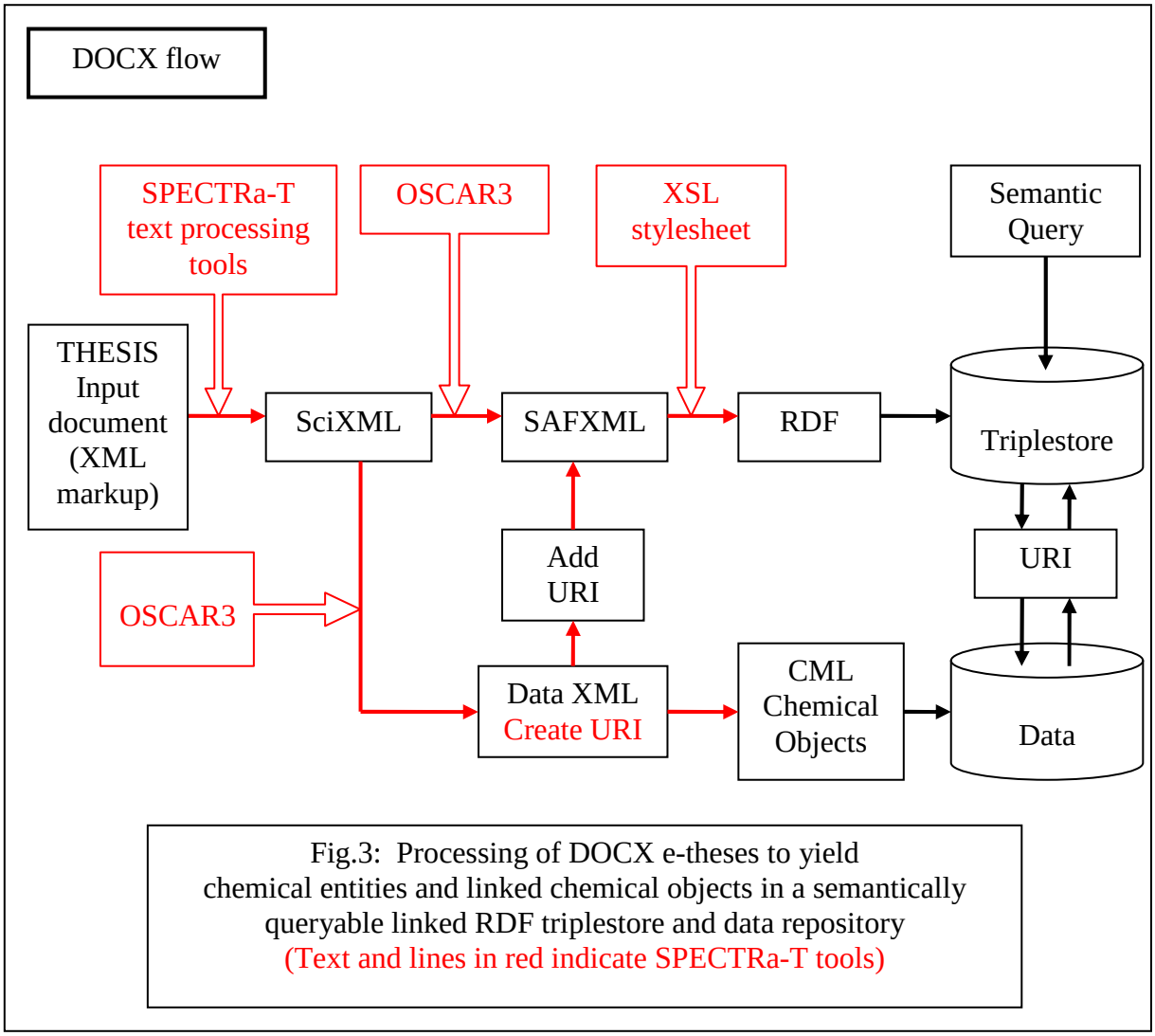

Fig.3: Processing of DOCX e-theses to yield chemical entities and linked chemical objects in a semantically queryable linked RDF triplestore and data repository
(Text and lines in red indicate SPECTRa-T tools)

Discussion

The SPECTRa-T project adopted a proof-of-concept approach to data-mining from chemistry e-theses. It focused on specific sub-disciplines within chemistry, utilising a text-mining tool designed for applications within the literature of those disciplines, and investigated only a limited range of thesis file formats. The techniques and results we have reported for selected chemistry sub-disciplines are thus not necessarily applicable to other areas of chemistry.

Nevertheless, despite this particular focus, there are important and much broader lessons that can be drawn from the project and applied to other sciences and indeed to the arts and humanities.

First, our results underline the need for institutions to determine what purpose theses might serve in each discipline beyond their immediate role in assessing a student's research. In the sciences theses should be regarded not simply as objects destined for preservation and archival status, but rather as unique resources containing potentially valuable data that must be made extractable and re-usable. This potential will not be realised by accident: it requires those authorities with responsibility for overseeing the writing, submission and subsequent management of theses to implement appropriate strategies, applying both institution-wide and discipline-specific policies, guidelines and practices designed to ensure optimal conditions for re-use. In other words, institutional regulations should ensure that each thesis is fit for purpose. Even where mandatory regulations do not require it, thesis authors should be encouraged to regard re-use of their work as a logical and desirable outcome, and to prepare accordingly. (We recognise that there are circumstances, such as confidentiality of data for security or commercial reasons, where authors will not wish their work to be publicly accessible or re-usable.)

Secondly, the project has highlighted specific problems with files in PDF format when text-mining techniques are applied. While PDF is currently the *de facto* standard for theses deposited in institutional repositories, we have shown that it is significantly less well-suited to text-mining than is marked-up text. PDF/A[6], which was developed specifically to aid the preservation of a file by tagging the document's semantics, is still essentially a page description format in which continuous text may be broken by images, tables, or embedded objects (all of which are important elements in chemistry e-theses) and thus fails to provide content in a contextual form amenable to text-mining tools. For data mining in chemistry, a marked-up document file format (such as Office Open XML) for deposited e-theses should be available in preference to PDF. Informal observation indicates that most e-theses are drafted in Word before conversion to PDF takes place as part of the submission process. It would thus be a relatively simple matter to require deposit of the Word version of the thesis, and then to convert it to XML as a standard process that would render it far more amenable to text-mining.

This raises an important question about the role of the institutional repository service: is it to be primarily an archive, ensuring that theses are carefully curated in protective custody; or should it undertake a more proactive role in promoting their re-use, helping to develop the tools and services that this would require? Or to put it in more immediate terms, should it plan to acquire and manage both PDF and XML versions of the thesis, with each format serving a specific purpose? With such a development the PDF file would be held in a managed preservation environment, while the XML file would be made openly accessible as the re-usable object. It should be noted that XML files are themselves increasingly regarded

as a preservation-friendly format. Betsy Fanning concluded her recent Technology Watch report[7]:

"It is important to note that one file format may not fit all needs in an organisation. As the organisation's needs change, the file formats chosen need to be reviewed to ensure that they fill the needs and meet the ever changing regulatory compliance requirements. The development of file formats such as XPS and OOXML should be followed and implemented where appropriate in an organisation. Increasingly, file formats are being based on XML which provides a level of interoperability amongst the file formats and aids in the preservation of electronic documents created using the formats."

If both PDF and XML files are to co-exist as suggested, the organizational structure of the repository service might also need to be reassessed, as the two files need not necessarily be held in the same repository. In our earlier SPECTRa project[8] we postulated that institutions might develop a federated repository structure in which departmentally-managed repositories (envisaged by SPECTRa as the front line in capturing experimental data from laboratory research) were linked with a central archival repository. Similarly, it is arguable that discipline-specific text-mining tools, together with the triplestore and data repositories into which text-mined outputs are to be deposited, will best be managed by subject experts operating within the same disciplinary environment.

The implications for institutional repository managers and librarians go further. As SPECTRa-T has demonstrated, much of the effort involved in developing text-mining tools and workflows is unavoidably specific to individual disciplines. It follows that those who are responsible for designing and delivering institutional repository services need to work closely with researchers in different subjects, in order to understand their needs and co-ordinate departmental and institutional activities.

We have also identified continuing uncertainties regarding IPR in data. It is clear that the ownership of data and any associated rights has important implications for data management in support of knowledge discovery. While facts cannot be copyrighted, the position with regard to derived data and databases is more complex[9], and the interests of the individual researcher and of the parent institution may not be the same. Legal advice - or repository managers' caution in interpreting it - has thus far had a restraining influence that favours the institution over the individual. Quality data are a major, but under-valued, asset, and university authorities need to appreciate this if they are to be persuaded to adopt policies that will foster data re-use. We believe that SPECTRa-T's work, by demonstrating the feasibility of extracting and exposing thesis-derived data to semantic querying, will help to persuade institutions that they can best realise the value of these assets by encouraging their discovery.

The IPR issues are equally important outside the host institution. There is increasing international support for the principles of Open Data, which require that all data produced by publicly-funded research should be made freely accessible and re-usable. To facilitate this, the ownership of scientific data and licensing arrangements for data re-use need clear guidelines that can be adopted uniformly across the research community to ensure consistent practice.

Conclusions

We have shown that:
- current practice in prescribing PDF as the preferred document format for deposition of chemistry e-theses discriminates against the possibility of data-mining;
- Named Chemical Entities and Chemical Objects (*e.g.* molecules, spectra) can be extracted routinely and automatically in high volumes from e-theses, transformed into metadata and deposited into data repositories and triplestores;
- RDF-based semantic querying of the data extracted from chemistry theses is a realisable goal.

References

1. SPECTRa-T (Submission, Preservation, and Exposure of Chemistry Teaching and Research Data from Theses) http://www.lib.cam.ac.uk/spectra-t/
2. Office Open XML.: Microsoft originally developed the specification as a successor to its binary Microsoft Office file formats and it was handed over to ECMA International to be developed as the ECMA 376 standard, which was published in December 2006. (http://en.wikipedia.org/wiki/Office_Open_XML)
3. SciBORG Project http://ww.sciborg.org/
4. Corbett P., Murray-Rust P.: *High-Throughput Identification of Chemistry in Life Science Texts*, Computational Life Sciences II, pp107-118 (Springer, Berlin, 2006)
5. (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.
6. ISO 19005-1:2005, Document management — Electronic document file format for long-term preservation — Part 1: Use of PDF 1.4 (PDF/A-1) http://www.aiim.org/documents/standards/19005-1_FAQ.pdf
7. Fanning B.A.: Preserving the Data Explosion: Using PDF (Digital Preservation Coalition & AIIM, April 2008) http://www.dpconline.org/docs/reports/dpctw08-02.pdf
8. SPECTRa (Submission, Preservation, and Exposure of Chemistry Teaching and Research Data) http://www.lib.cam.ac.uk/spectra/
9. Korn N, Oppenheim C., Duncan C: IPR and Licensing Issues in Derived Data (JISC, May 2007) http://www.jisc.ac.uk/media/documents/projects/iprinderiveddatareport.pdf

Note: The final report of the SPECTRa-T project will be published on the project website http://www.library.cam.ac.uk/spectra-t/ later in 2008.

Acknowledgements

This project was funded by the Joint Information Systems Committee (JISC) under its Digital Repositories programme.

The project team are grateful for assistance provided by Peter Corbett (SciBorg Project, University of Cambridge).