


Cover Sheet for Proposals <i>(All sections must be completed)</i>			
Name of Tender:	Small-Scale OAI Object Re-Use and Exchange Experiments		
Name of Bidder:	University of Cambridge		
Name of Proposed Project:	TheOREM		
Name(s) of Project Partner(s):	University of Cambridge		
Full Contact Details for Primary Contact:			
<p> Name: Jim Downing Position: Software Officer Email: ojd20@cam.ac.uk Address: Unilever Cambridge Centre for Molecular Science Informatics, University Chemical Laboratory, Lensfield Rd, Cambridge. CB2 1EW Tel: 01223 336300 Fax: 01223 336362 </p>			
Length of Project:	6 months		
Project Start Date:	22/02/2008	Project End Date:	22/08/2008
Total Funding Requested from JISC:	£29,324		
Outline Project Description			
<p> TheOREM (Theses with ORE Metadata) aims to: - <ul style="list-style-type: none"> ● Test the applicability of the ORE standard in a realistic scholarly setting - thesis description, submission and publication. ● Demonstrate the advantages of the ORE approach in complex object publication, by combining it with existing web-standards compliant technologies. ● Provide examples to fully exercise the ORE specifications in order to provide validation and future direction. </p> <p> TheOREM will contain two main strands. Firstly, we will create a small corpus of ideal born-digital theses based on real theses and describe these as completely as possible using ORE. Secondly, we will define a realistic scholarly scenario in which such theses might be handled, and implement demonstrators for each component system in the scenario in order to show the capabilities and limitations of ORE. </p>			

Summary

1. This proposal is submitted by the University of Cambridge, seeking funds from the JISC for project TheOREM, an experiment into the applicability of the Open Archives Initiative Object Reuse and Exchange[8] (ORE) standard in the description and publication of rich digital chemistry thesis objects.
2. The project will apply ORE to fully describe and enhance a small corpus of openly licensed theses such that they represent the ideal digital thesis, and then build a demonstrator system around a thesis submission scenario to validate and challenge the ORE approach and technology.
3. The project plan and the results of any initial investigation will be presented at the ORE launch meeting on March 3rd 2008 and progress will be reported at the ORE day at Open Repositories 2008. The project will be reported formally in a final report and presented at an appropriate conference.
4. The project will begin on 22nd February 2008 and run until 22nd August 2008.

Background

5. There is a major effort worldwide to capture student theses as born-digital objects. Much of the effort to date has concentrated on PDF documents as a proxy for the printed version, but PDF cannot achieve the full potential value of born-digital theses. Publication of properly structured theses can allow rapid exploration of material, for example down to individual diagrams or the sub-subsection level. Individual citations can be extracted automatically. Supporting data can be associated with the thesis directly rather than reduced to tabular or graphical form in appendices. Future readers will want not only the document level descriptive metadata (author, institution, title, keywords, dates) but also access to subcomponents (data, tables, references, diagrams, etc.). For example a reader might wish to survey all the analytical instrumentation used within the last year and therefore only need access to the "Materials and Methods" sections, but to wish this for every thesis. Similarly a bibliographer might wish to analyse communities of practice from citations. What proportion of references are identical to those in earlier theses from the same institution? Longitudinal studies may be possible - retrieve all images of protein gels over the last 10 years and see whether the quality has increased.
6. Current package-based approaches to the transfer and publication of complex digital objects result in problems of balkanisation due to the multiplicity of available packaging standards that complicate interoperability between systems and the discovery and reuse of complex object components. They usually use a "pass-by-value" paradigm of data transfer, which creates additional complexities in data duplication, trust / access control and revisioning. The ORE "pass-by-reference" approach allows a much more flexible, disaggregated approach to complex object description, access and transfer.
7. There is a growing realisation in the repository community that the web is not only an essential part of content delivery, but also an underused and undervalued architectural substrate for the repository ecosystem that offers easier interoperability between systems within our domain, and with the web community at large. ORE works within the constraints of the web architecture, therefore it will be possible to combine it with other web standards (e.g. Atom Publishing Protocol[1], SWORD[13], XACML[15]) to integrate large scale infrastructures.

Project Aims

8. The general aims of the project are as follows: -
 1. Test the applicability of the ORE standard in a realistic scholarly setting — thesis description, submission and publication.
 2. Demonstrate the advantages of the ORE approach in complex object publication, by combining it with existing web-standards compliant technologies.
 3. Provide examples to fully exercise the ORE specifications in order to provide validation and future direction.

Project Description

9. The experimentation in TheOREM will contain two main strands. Firstly, we will create a small corpus of ideal born-digital theses based on real theses and describe these as completely as possible using ORE. Secondly, we will define a realistic scholarly scenario in which such theses might be handled, and implement demonstrators for each component system in the scenario in order to show the capabilities and limitations of ORE.

ORE Description of Chemistry Theses

10. The first phase of the project involves improving a small (≈ 10) corpus of theses such that the data is extracted and groomed (quality checked and improved), the source versions of each chapter are available separately from the printable PDF and so on. This aggregation of digital objects will be described in separate metadata files using, in particular, the ORE data model and the Scholarly Works Application Profile (SWAP[12]). In the course of this work, we will investigate methods of extending ORE to add chemistry-specific and thesis-specific semantics to ORE Resource Maps, investigating how ORE and SWAP interact and investigating the implications of using various serialization formats for ORE Resource Maps.

11. As part of the JISC-funded SPECTRa-T project (still running at time of writing), we collected a number of chemistry theses that were available under an open license. A subset of these will form the corpus for TheOREM, allowing our ongoing progress and final results to be openly published.

12. The Resource Maps created will contain: –

- Structural metadata — defining the relationships between the PDF version of the thesis, source versions of each chapter, appendices, supporting data.
- Descriptive metadata — basic Dublin Core metadata, citations, embargo details, license details, ontology links.

Thesis Submission Scenario

13. The thesis submission scenario is depicted in figure 1, below.

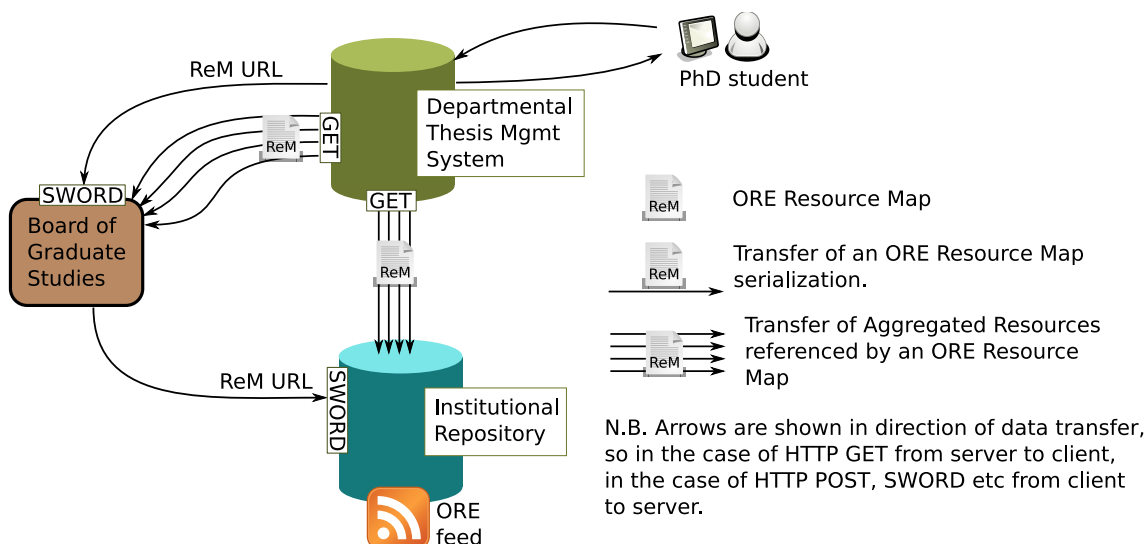


Figure 1: Schematic of the TheOREM submission scenario

14. The PhD candidate writes their thesis. We assume the existence of systems provided at department level to support the collection and management of experimental and processed data (“Departmental Thesis Management System” in figure 1). This is a reasonable assumption; such systems are currently being developed (e.g. Imperial College Computational Chemistry system from the SPECTRa project[10]). Continuing this work to create a fully

working system is beyond the scope of this project — TheOREM will create a demonstrator thesis management system with working interfaces as required by this scenario.

15. On submission of the thesis, the thesis management system submits the ORE Resource Map representing the thesis to the Institution registry / board of graduate studies using a modification of the SWORD protocol for ORE. It is important to freeze this snapshot of the thesis, so it is likely that the registry will dereference and store all the components of the thesis. An alternative that could be investigated if time is available would be for the thesis management to support revisioning.

16. The thesis will almost invariably require corrections after the PhD is examined - made through the thesis management system. The student adds metadata required to control the embargo of parts of the thesis on publication (such as those discussed in the SPECTRa project[10]).

17. When the PhD is awarded, the Resource Map is submitted (again using a modified SWORD interface) to the institutional repository (IR). The IR will dereference and store the components of the thesis, issue persistent identifiers for all the resources, apply access controls consistent with the embargo metadata and make the ORE Resource Map discoverable through an Atom Syndication[2] feed.

18. TheOREM will implement demonstrator software systems that show these interactions working, and that allow them to be initiated through a web site interface.

19. TheOREM will use the ORE enhanced resources published by the demonstrator IR feeds in data mashups with cited sources (e.g. JSTOR[7]) and other publically available sources of chemistry data to provide examples of the value of the ORE approach.

Project Management

Plan of Work and Deliverables

Months	Activity	Deliverables
1	Detailed project planning	Project plan, early dissemination presentations at ORE meeting in March and Open Repositories conference in April.
1 - 2	Preparation of theses and ORE description	Method for extending ORE to describe chemistry theses, corpus of sample complex thesis objects.
3 - 6	Development of demonstrator systems to implement the submission scenario	Demonstrator systems, revised ORE model and methodology.
6	Reporting and final dissemination	Final report, dissemination through relevant CRIG event.

Risk Management

Risk	Mitigation
Staff recruitment and retention	Project will be performed by existing staff (no risk in recruitment). Technical skills required are held by other members of research group — short periods of project staff loss should not have a significant impact on the project.
Changes to standards (ORE, SWORD etc) during project	Project team will maintain close links with committees of all relevant standards that are still under development.
Technical Challenges	Any technical challenges are a relevant outcome from this investigation; if technical limitations of ORE prevent us from achieving the project deliverables, the reasons will be fully reported.

Engagement and Dissemination

20. We are involved in ORE, both in an advisory capacity (on the advisory board) and involved in technical discussions regarding the specification.

Any uncopyrightable material (data) will be placed in the public domain in order to promote the adoption of open community norms in chemistry. All other materials created (reports etc.) will be made available under a Creative Commons attribution license.

Project Team

27. Members of the project team have a wealth of experience and current involvement in related projects to TheOREM which will be beneficial to this project.

Dr Peter Murray-Rust - Principal Investigator

28. 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[3], now used by groups such as the European Patents Office, and have a long involvement in applying semantic web techniques (from which ORE derives) to chemistry[14, 6]. Peter Murray-Rust was associate PI on the SPECTRa[10] and SPECTRa-T[11] projects, is on the ORE advisory board, and is PI on the Microsoft funded OREChem project[9].

Jim Downing - Project Manager & Technical Lead

29. Jim Downing is Software Development Officer at the Unilever Centre. He has 9 years experience in designing and developing knowledge management and information systems. Jim has been working in repositories in higher education for 4 years, having worked on the SPECTRa and SPECTRa-T projects. He is a committer for the DSpace Open Source software project, and is co-chair of the JISC Common Repositories Interoperability Group[4]. Jim was part of the JISC Deposit API working group, and advised the JISC SWORD[13] project.

Nick Day - Technical

30. Nick Day has been researching in the Murray-Rust group for over 3 years, and developed the CrystalEye[5] crystallography repository system as part of his PhD, after gaining a MChem(hons) in Chemistry from Oxford University. He has strong software development and semantic web skills.

References

- [1] Atom Publishing Protocol — <http://www.rfc-editor.org/rfc/rfc5023.txt>
- [2] Atom Syndication Format — <http://tools.ietf.org/html/rfc4287>
- [3] Chemical Markup Language (CML) — <http://www.xml-cml.org/>, <http://sourceforge.net/projects/cml/>
- [4] Common Repositories Interoperability working Group — <http://www.ukoln.ac.uk/repositories/digirep/index/CRIG>
- [5] CrystalEye — <http://wwmm.ch.cam.ac.uk/crystaleye/>
- [6] Golem, an ontology language and development toolkit for CML — <http://wwmm.ch.cam.ac.uk/wikis/wwmm/index.php/Golem>
- [7] JSTOR - The scholarly journal archive — <http://www.jstor.org/>
- [8] Open Archive Initiative Object Reuse and Exchange standards — <http://www.openarchives.org/ore/>
- [9] OREChem, a Microsoft sponsored collaboration between Open Archives Initiative, University of Cambridge, Soton University, University of Indiana and Penn State University — <http://www.orechem.org/>, <http://sourceforge.net/projects/orechem>
- [10] SPECTRa - Submission Preservation and Exposure of Chemistry Teaching and Research data. A JISC-funded project. Final Report — A. P. Tonge and P. B. Morgan, <http://www.lib.cam.ac.uk/spectra/FinalReport.html>
- [11] SPECTRa-T — <http://www.lib.cam.ac.uk/spectra-t/>

- [12] Scholarly Works Application Profile. A Dublin Core Application Profile for describing scholarly works held in institutional repositories — <http://www.ukoln.ac.uk/repositories/digirep/index/Eprints.Application.profile>
- [13] SWORD - Simple Webservice Offering Repository Deposit — <http://www.ukoln.ac.uk/repositories/digirep/index/SWORD>
- [14] P. Murray-Rust, and H. S. Rzepa "Towards the Chemical Semantic Web", Proc. 2002 International Chemical Information Conference,, ed H. Collier. (Infonortics) 2002, pp 127-139.
- [15] XACML - XML Access Control Markup Language — http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

