ICE-Theorem - End to end semantically aware eResearch infrastructure for theses

Peter Sefton
sefton@usq.edu.au
University of Southern Queensland

Jim Downing
ojd20@cam.ac.uk
Nick Day
ned24@cam.ac.uk
University of Cambridge

OpenRepositories 2009, Atlanta, Georgia USA
2009-05-19

Table of Contents

Abstract:

ICE-TheOREM was a project which made several important contributions to the repository domain, promoting deposit by integrating the repository with authoring workflows and enhancing open access by prototyping new infrastructure to allow fine-grained embargo management within an institution without impacting on existing open access repository infrastructure.

In the area of scholarly communications workflows, the project produced a complete end-to-end demonstration of eScholarship for word processor users, with tools for authoring, managing and disseminating semantically-rich thesis documents fully integrated with supporting data. This work is focused on theses, as it is well understood that early career researchers are the most likely to lead the

charge in new innovations in scholarly publishing and dissemination models.

The authoring tools are built on the ICE content management system, which allows authors to work within a word processing system (as most authors do) with easy-to-use toolbars to structure and format their documents. The ICE system manages both small data files and links to larger data sets. The result is research publications which are available not just as paper-ready PDF files but as fully interactive semantically aware web documents which can be disseminated via repository software such as ePrints, DSpace and Fedora as complete supported web-native **and** PDF publications.

ICE-TheOREM combined the Object Reuse and Exchange (OAI-ORE, IONSREPORT 2008) and SWORD-APP protocols to transfer content between a content management system, a thesis management system and multiple repository software packages and looked at ways to describe aggregate objects which include both data and documents, and to represent structure withing thesis documents. This can be generalized to domains other than chemistry. ICE-TheOREM has demonstrated how focusing on the use of the web architecture (including ORE) enables repository functions to be distributed between systems for complex, data-rich compound objects.

1 Introduction

The TheOREM project exercised the OAI-ORE protocol in the context of chemical theses by modeling the thesis as an aggregate of chapters and supporting information, and by proposing mechanisms to leverage ORE in a hypothetical scenario describing a thesis submission and consequent deposit and publication in an Open Access Institutional Repository. The ICE (Integrated Content Environment) (Sefton 2006) extension to that project showed how chemical theses could be authored in a word processing environment, following from proof of concept work presented at the Electronic Theses and Dissertations conference in 2007 (Murray-Rust 2007). We have been able to demonstrate theses that are both 'supported' by data in Neylon's terms (Neylon 2008) and are datuments (Murray-Rust & Rzepa 2004) in that they are hypertext aggregations of document and data, which are both human and machine-readable.

ICE-TheOREM was a joint project between the University of Cambridge (UC) and the University of Southern Queensland (USQ) funded by the JISC. At USQ, there was a team involved in this work: Oliver Lucido, Ron Ward, Linda Octalina, Bronwyn Chandler and Duncan Dickinson all assisted in programming and project management. At Cambridge, Nick Day was the implementer, with support from Joe Townsend.

The technical motivation for the project was to test the OAI-ORE standard, asking:

- Is it applicable and useful for managing thesis workflows and the relationships between documents and data?
- What are the different ways of using ORE in this context?
- How might the SWORD-APP (Simple Webservice Offering Repository Deposit AtomPub profile) (Downing et al. 2009) and ORE combine?

There are various approaches for using ORE with the SWORD-APP protocol to move theses and linked-data from one system (the client in the context of this transfer) and a repository system (the server), and the most practical option turned out to be one which used AtomPub to transport ORE Resource Maps (ReM), and consequently did not to use SWORD's support for content packaging.

The project designed a thesis process based around a purely web architecture, looking for interoperability between systems and, looking towards a repository modeled as a set of services; an idea promoted by repository theorists, but which is lacking in most IR installations, which tend to be monolithic.

From the point of view of open access, we tested whether disaggregation of theses into chapters could promote open access, by making it easier to disseminate open parts via existing OA systems and hold-back embargoed content in an upstream system. In this model we posited that promoting embargo of sensitive chapters could accelerate publication of the remainder of the thesis. To explore this we had to look at issues such as:

- How can embargo metadata be passed between systems?
- When should it be created?
- Where and how should it be stored?
- How can candidates be tracked once they've graduated?

Answers to these questions emerge in the project narrative below, which is organized to follow the life-cycle of a thesis.

In this paper we follow the life-cycle of a thesis starting with writing and supervision, then examination and deposit of a thesis showing where the ICE-TheOREM project (Jacobs 2008) has produced proof of concept innovations that promise to improve on current repository practice. While the project was exploratory in nature there have been some concrete outcomes.

2 Thesis life-cycle stages

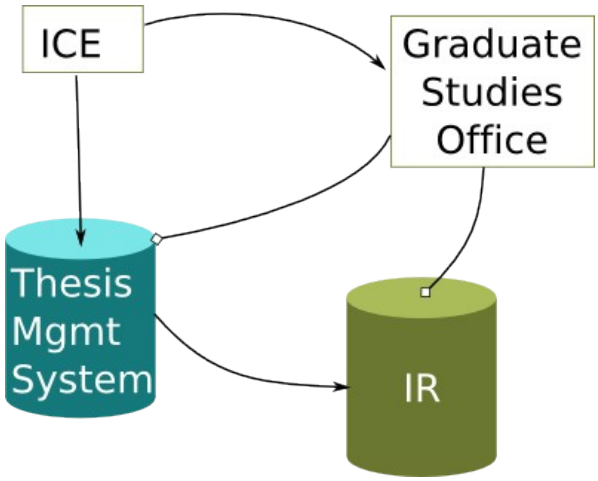


Illustration 1: Overall thesis workflow with thesis repository and Board of Graduate Studies (BoGS)

The thesis is authored in the ICE system, which also handles the interactions between candidate and supervisor through annotations. When the thesis is submitted, a copy is sent to the Thesis Management System (TMS) and also to a hypothetical system representing the Graduate Studies Office (GSO), where it is sent or made available to the examiners. We did not prototype mechanisms for distributing copies to examiners or to model the examination and correction process, but chose to focus on the interactions between the systems shown. When all necessary corrections have been made and an updated version has been transferred to the TMS, GSO sends a message to the TMS to make the thesis available to the Institutional Repository (IR), and another to the IR that it should start to collect and republish the thesis.

This design was motivated by a desire to prototype a system that could work without an institutional mandate for thesis deposit or Open Access, and that decentralized control over embargo. The latter is important as it is supervisors or candidates, rather central university administration, that most often make decisions about embargo.

The centerpiece of the ICE-TheOREM project has been a Master's thesis by Malcolm Tait. This thesis was collected as part of the JISC sponsored SPECTRa-T project (Murray-Rust 2007) in its source format (Microsoft Word .doc) with permission to process and republish. It is a typical thesis in the area of chemical synthesis, with a review of the properties of the molecules in question and previous work, followed by a discussion of the work conducted. The narrative sections of the document are interspersed with tabular data and diagrammatic representations of molecular structures, and the the appendices contain large amounts of detailed procedure description and characterization data. The thesis is shown here in the ICE system, running in the virtual machine we created for the project (http://hdl.handle.net/102.100.100/32). The thesis is broken up into multiple source documents, one for each chapter or section of the work, in this case in Microsoft Word (.doc) format, but OpenOffice.org (.odt) files are also supported. The view shown here is a web-rendered view of the thesis, ICE converts each part into HTML, and also creates a PDF version.

2.1 Authoring

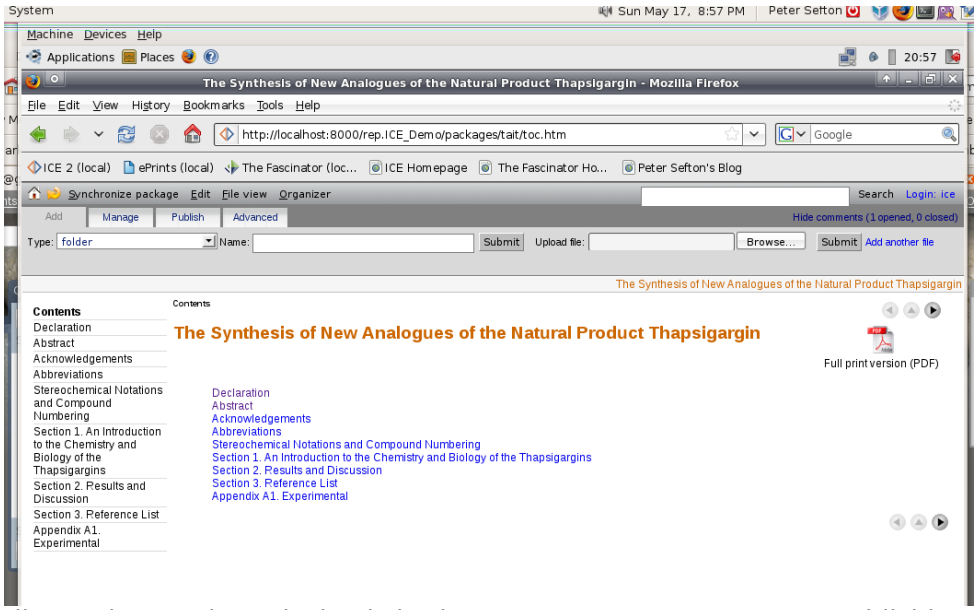


Illustration 2: The Tait thesis in the ICE content management/publishing system

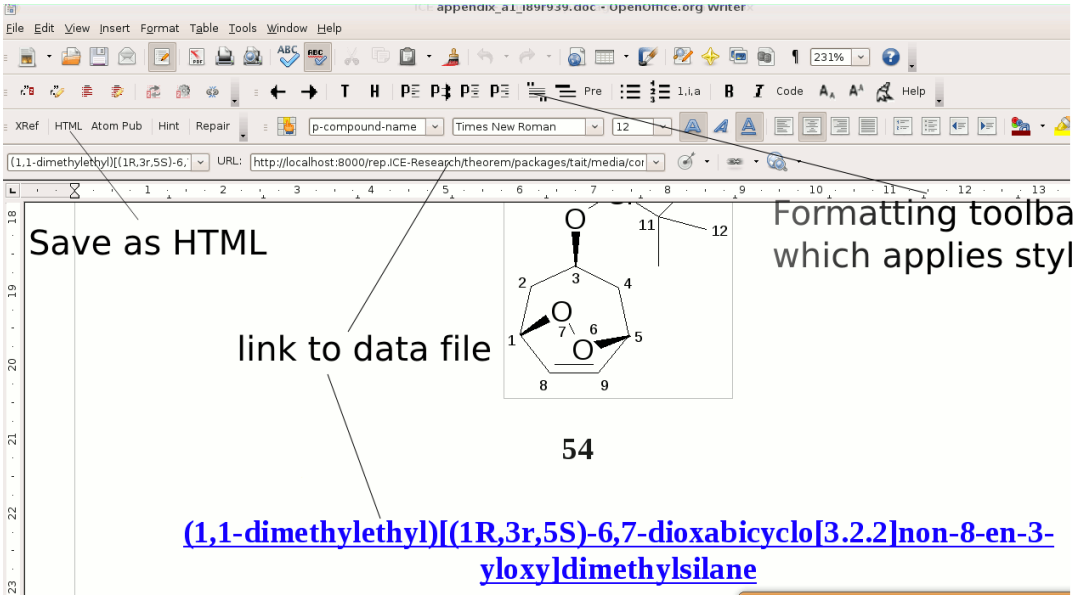


Illustration 3: Editing a document

The key features of an ICE document are highlighted in Illustration 3. It uses styles to convey structural information about a document, the author applies styles using a toolbar, and the document can be converted to HTML format or sent to a website (usually a weblog) via the Atom Publishing protocol. In this case, though, the author does not have to click any buttons in the word processor to see the thesis in HTML, they look at it through the ICE web application, which runs on their desktop – changes to the document are automatically reflected in the web-view when the author refreshes the page. This is an important feedback mechanism which helps to improve the quality of documents created in ICE – any inconsistencies between the print and web view can be spotted by the author immediately. This contrasts with workflows where authors send documents away for processing and may not see the results for hours, days or months.

One of the key features developed for this project is the ability to link to data in a meaningful way. In this case, the image of a molecule, `(1,1-dimethylethyl)[(1R,3r,5S)-6,7-dioxabicyclo[3.2.2]non-8-en-3-yloxy]dimethylsilane` is linked to a Chemical Markup Language file describing it so the ICE application embeds a 3D rendition of the molecule of the page in its web-rendition. This data accompanies the document throughout its life-cycle as it move through the workflow described in this paper.

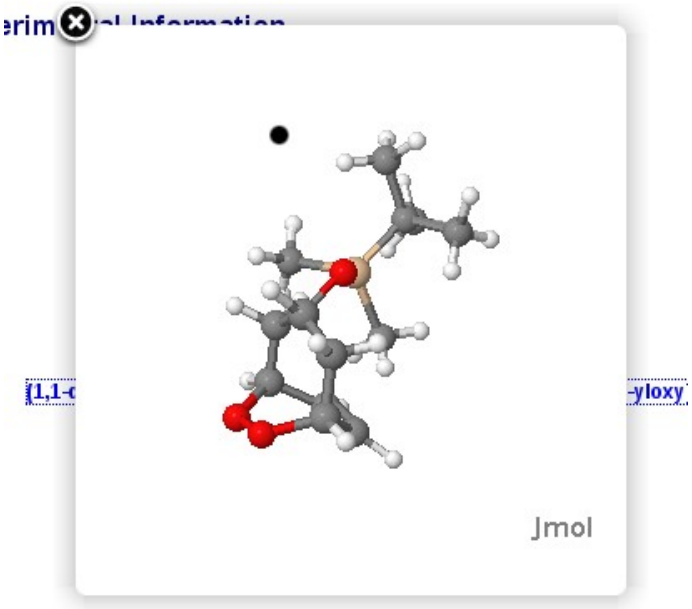


Illustration 4: Interactive data-aware documents

The ICE system allows for stand-off annotation of documents in a way that is similar to the digress.it tool, formerly commentpress. Supervisor(s) and peers are able to comment on a document without changing it.

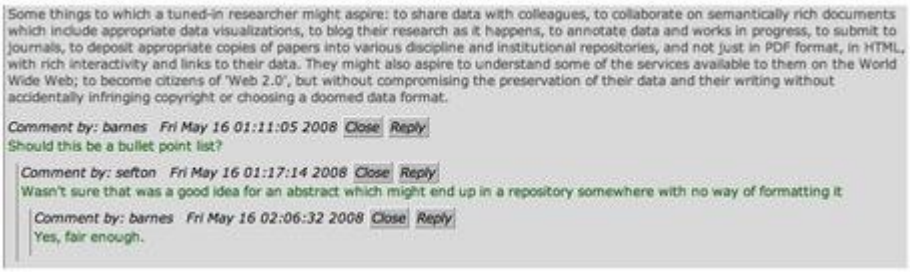


Illustration 5: Annotation

2.2 Submission

When the document is ready for submission for examination, the ICE-TheOREM model proposes a repository which belongs to the graduate studies office, so the thesis needs to be deposited in that repository. This could be accomplished using a 'pull' process where the repository watches the ICE system and fetches theses with a certain flag set, such as *ready_ for_examination,* as described and prototyped in the competition entry for Open Repositories 2008 *Zero Click Ingest (Monus et al. 2008)*. In ICE-TheOREM we have used a push system, where the candidate uses the SWORD function to send the thesis to a thesis repository. This SWORD button is now in use at USQ with the ePrints institutional repository as well, allowing authors to post completed works as soon as they have been accepted into a journal or delivered at a conference, such as an earlier version this paper which is available in HTML as well as PDF (Sefton et al. 2009).

The use of SWORD here is special – we are using SWORD as a transport but OAI-ORE as well, to describe the structure of the thesis as an aggregate object.

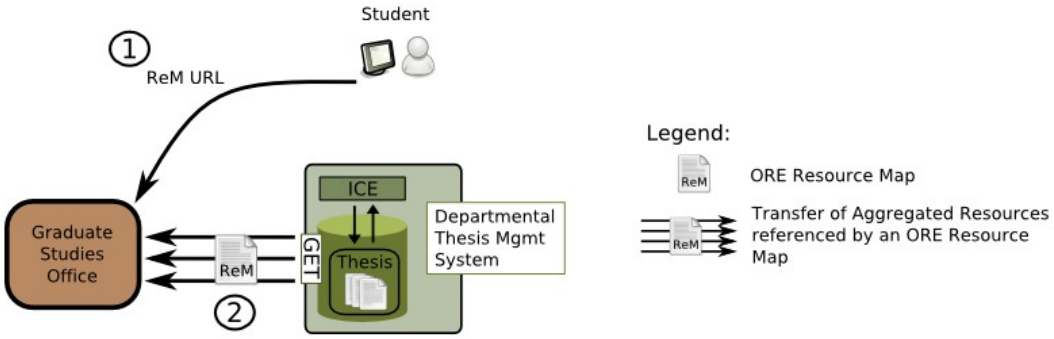


Illustration 6: Initial thesis submission - schematic

But before we look at the submission process we need to consider embargo. One of the major contributions of ICE-TheOREM is a model for granular thesis embargo, allowing individual chapters or sections to be placed under embargo. While it is likely that this will be used for reasons of commercial exploitation or privacy of research subjects, there are few safe assumptions here; we have heard of a case where a PhD graduate was happy for an entire thesis to be made open access apart from the acknowledgements section. To model this situation, in our demonstration the acknowledgements section is placed under embargo.

Embargo metadata is encoded in a style:

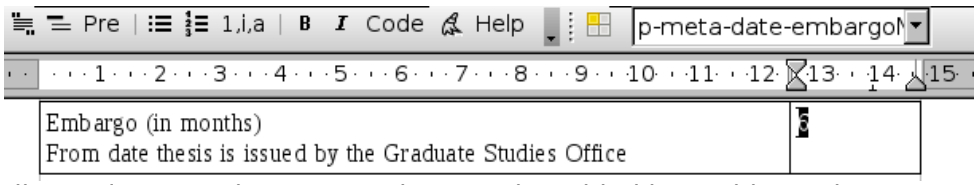


Illustration 7: Embargo metadata can be added in a table, and identified by using word processing styles

And ICE can extract the metadata:

```
<oai_dc:dc>

<dc:title>Acknowledgements</dc:title>

<dc:relation>date-embargoMonths::6</dc:relation>

</oai_dc:dc>
```

Illustration 8: Metadata extracted from the doucment by ICE

The initial demonstration encodes embargo information using a style, using a technique developed in the ICE and ICE-TheOREM projects.

When the thesis is sent to the thesis repository via SWORD, then the metadata is sent with it. We propose that the graduate studies office get the student to submit and validate an OpenId (Recordon & Reed 2006)– allowing the student to authenticate to administer embargoes after their institutional login expires by

authenticating with OpenID. While ICE and the thesis repository based on The Fascinator can both accept OpenId login, the details of managing student identity have not been worked out. Whilst the idea of using OpenId is an interesting possibility, it should be noted that embargoes that lift automatically after a fixed period of time avoid such technical and / or administrative overheads, and are likely to be preferable where thesis deposit is mandated.

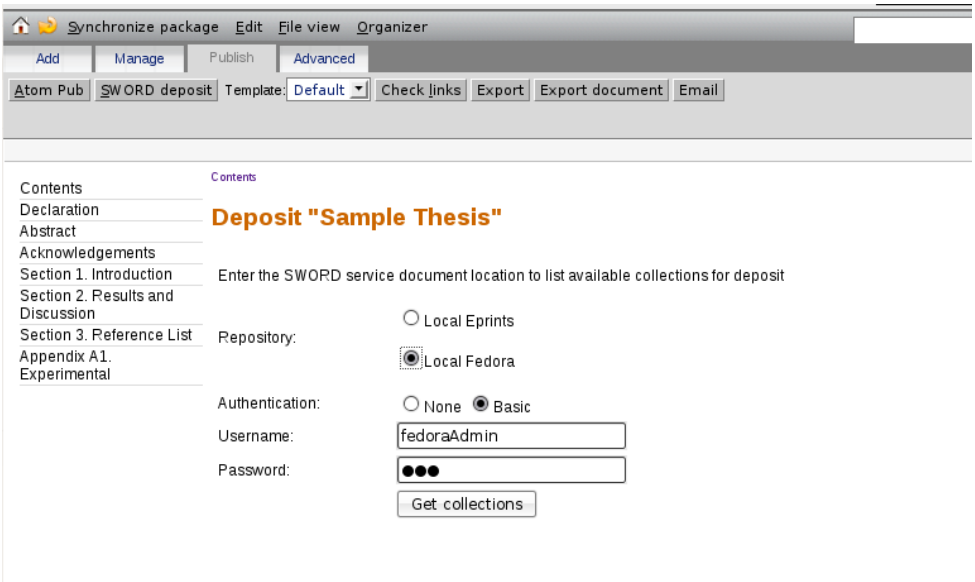


Illustration 9: SWORD deposit

The SWORD deposit contains an OAI-ORE payload.

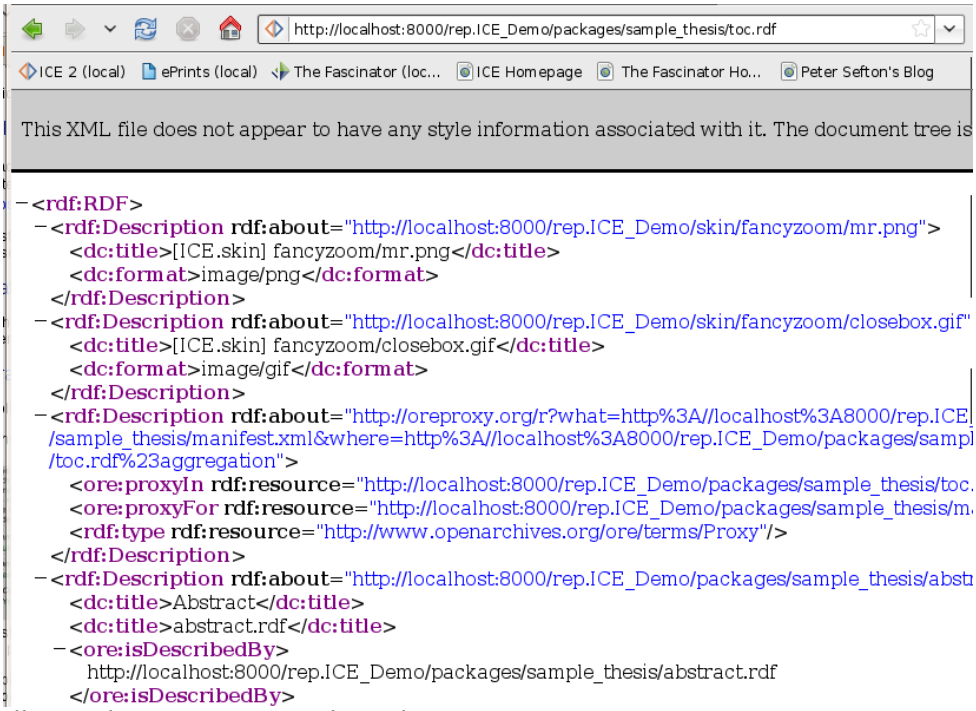


Illustration 10: SWORD deposit uses an ORE Resource Map

This XML is expressing the structure of the thesis.

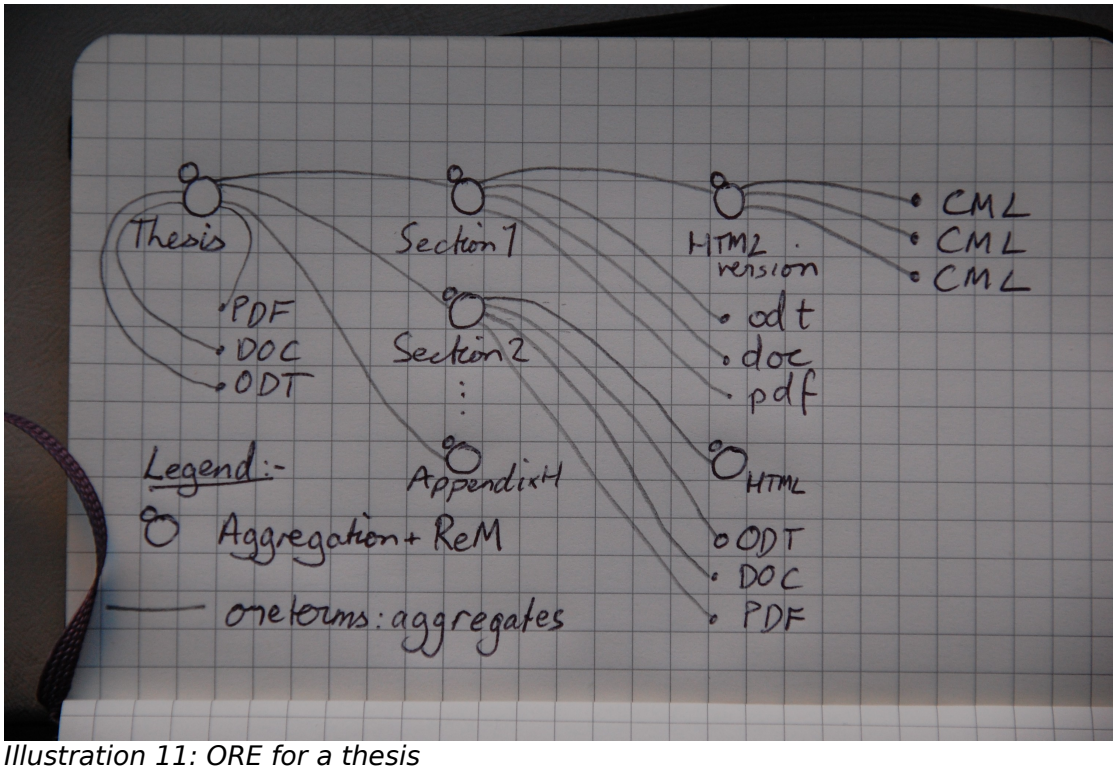


Illustration 11: ORE for a thesis

2.3 Importance of ORE

ORE is important to this work because it:

- Allows description of aggregate objects like theses.
- Can specify the relationship between two renditions of the same thing, such as HTML and PDF for a chapter.
- Can include external things like data files as part of an object.

(Currently repositories such as ePrints and DSpace do not do this at all well, content models for repository items are usually implicit.)

The use of ORE to describe a thesis as an aggregate of objects makes it easier to implement fine-grained embargoing, as in illustrations 12 and 13, which show the ICE-TheOREM mock-up repository, implemented using The Fascinator (Sefton & Lucido 2009) to serve theses from a Fedora 3 repository. Further discussion of the mechanics of using ORE to implement embargoed transfer can be found in Section 2.5 , below.

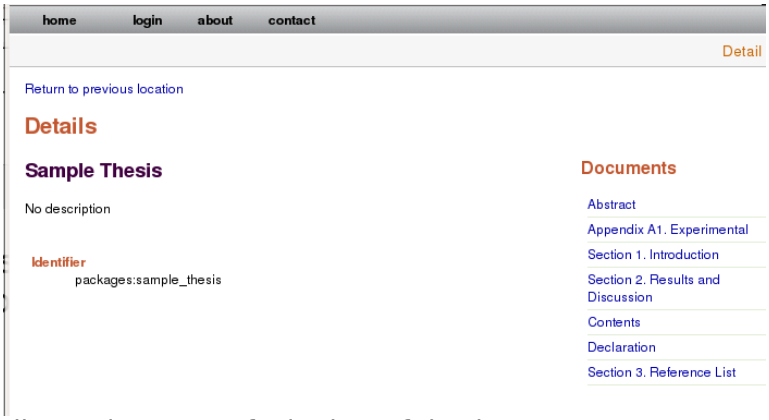


Illustration 12: Default view of thesis - no acknowledgements

Whereas if an administrator is logged in then the acknowledgements are visible.

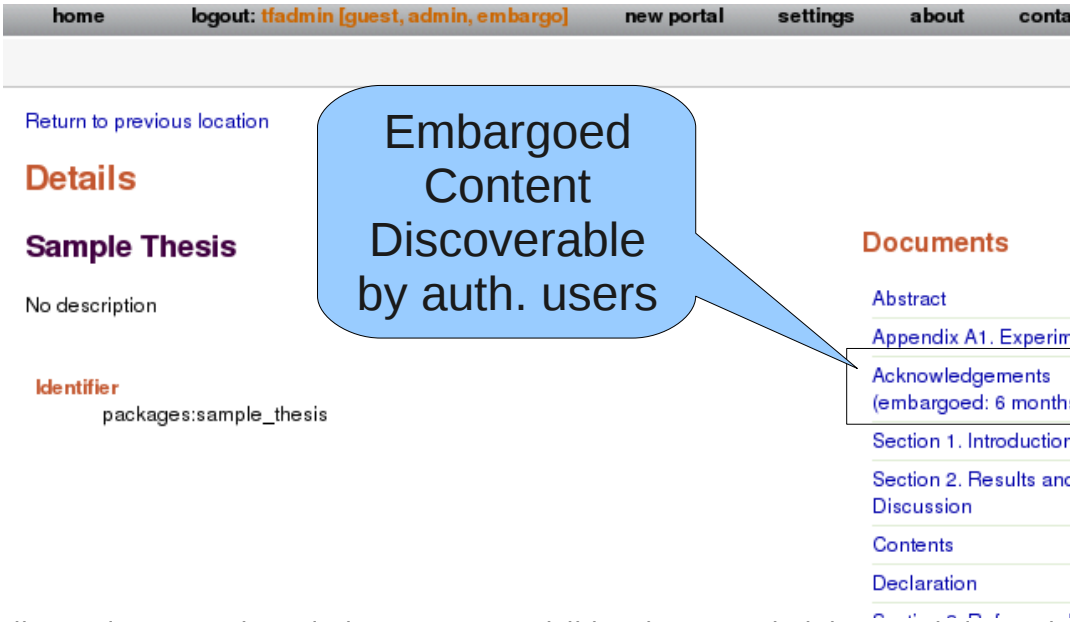


Illustration 13: Acknowledgements are visible when an administrator is logged in

The thesis repository is underdeveloped, with more work to do, but in a production version of the model presented here, the thesis repository would feed the institutional repository, using some approach to incremental embargo release (see Section 2.5)

2.4 ORE + SWORD

There are various approaches for using ORE with the SWORD-APP protocol to move theses and linked-data from one system (the client in the context of this transfer) and a repository system (the server), distinguished primarily by the way they use the ORE Resource Map (ReM):

2.4.1 ReM as Manifest

In this approach the entire contents of the thesis, including a ReM in a file that acts as a manifest, are bundled into a single package archive file, such as a zip file or tape archive (tar) file, which is then transferred using SWORD-APP. The ReM includes relative URLs to refer to, and describe, the contents. This is the specific situation for which SWORD was designed. The main advantage is that the client does not need to act as a server as well. Because all of the aggregate object's parts are transferred in a single HTTP request, this approach provides transactional guarantees that are more difficult to implement in the other approaches. The disadvantages of this approach are that it precludes the benefits of pass-by-reference of the other approaches, and replicates content in a way that makes it difficult to track and resolve copies.

2.4.2 ReM as a Shopping List, ReM as Road Signs

In terms of the initial transaction, these approaches are identical; the ReM is transferred using AtomPub. The SWORD extensions to support content packages are consequently not used, and the SWORD profile extensions will only be useful if there is some other some other requirement for them (e.g. mediated submission using OnBehalfOf). The difference is that in the *Shopping List* approach the server expects to dereference all of the aggregated resources immediately in order to republish them at new URLs, whereas the *Road Signs* approach uses the original URLs as they are, dereferencing only when necessary to access content. An important discriminatory feature between the approaches is the effect on access control. Since the server in the *Shopping List* approach republishes resources, access control is performed independently by the two systems, probably without co-ordination. We used the *Shopping List* approach in ICE-TheOREM, as it was more appropriate to separate concerns by having the repository system deal with access control and embargo management (discussed later), leaving the Thesis Management System free of these concerns. In other situations it would be more appropriate for the originating system to have the sole duty for access control, indicating the *Road Sign* approach as a better fit.

2.5 Incremental Embargo Release

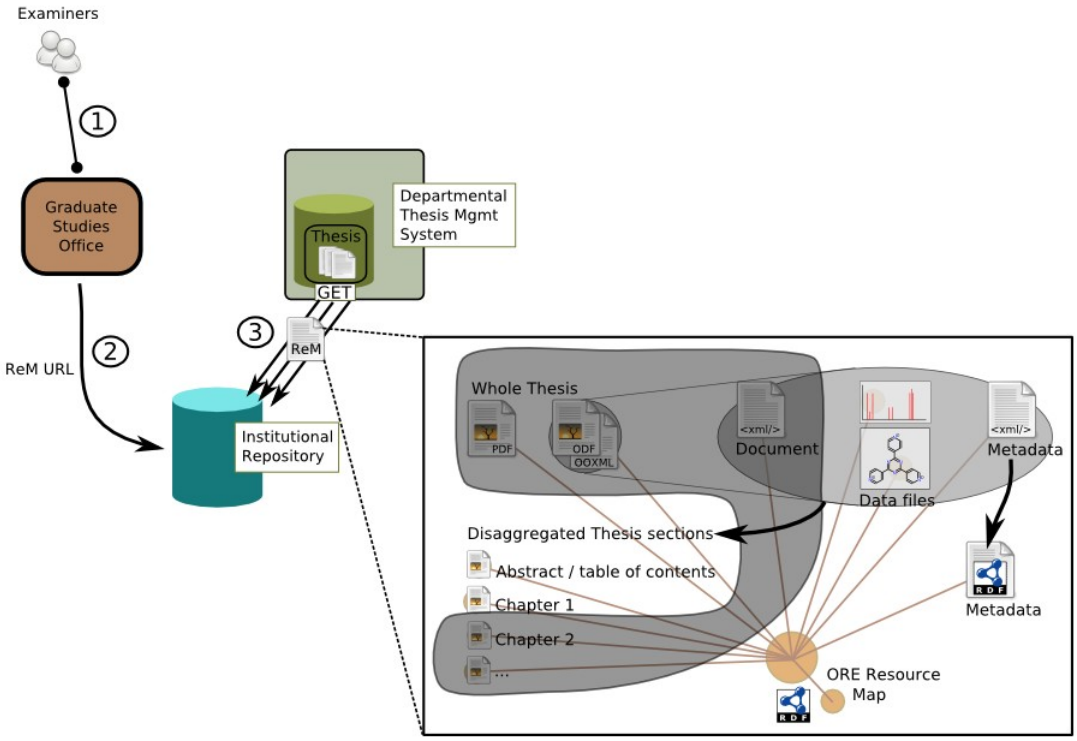


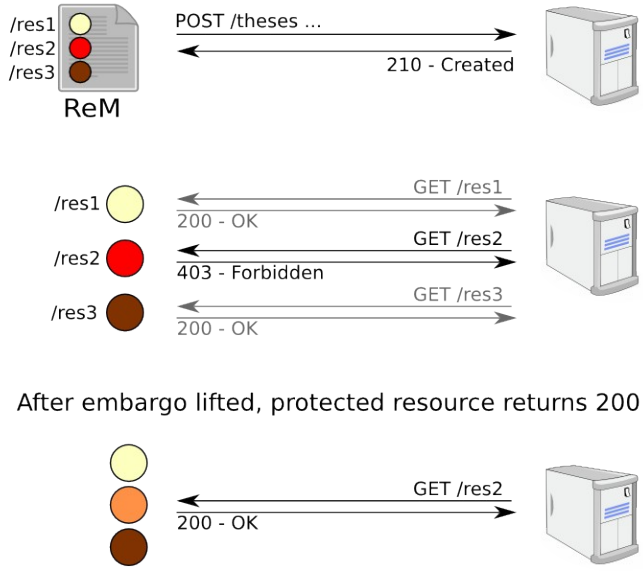
Illustration 14: The final stage – automated IR deposit

As indicated earlier, we were unable to fully implement a demonstration of incrementally transferring partially embargoed theses to an IR software, but here we propose three potential mechanisms that use standard web mechanism and ORE to implement the transfer of an aggregation as embargoes on parts of the aggregation lift over time. All three use SWORD + ORE to transfer the thesis ReM using the *Shopping List* recipe (described in Section 2.4.2), but have different approaches to incremental embargo release; pull (polling resource), pull (polling ReM), and (re-)push.

2.5.1 Scheme 1: Use HTTP Forbidden Status Code

In the first scheme (Illustration 15: Scheme 1) the thesis recipient is told of all the resources in the aggregation and polls each of them. It is up to the sending system to protect embargoed resources using HTTP authentication. This scheme requires the sending system to be able to act as a server as well as a client, and requires the recipient system to be able to authenticate with the sender. Continued requests for a resource that might stay embargoed for several years is also inefficient, although probably not problematically so since long (e.g. monthly) polling intervals would probably be acceptable.

Scheme 1: Send full ReM, recipient polls protected Resources



After embargo lifted, protected resource returns 200

Illustration 15: Scheme 1

2.5.2 Scheme 2: Pull Updated ReM

The second scheme (Illustration 16: Scheme 2) has the recipient system polling the ReM representing the thesis, and takes advantage of HTTP caching mechanisms (e.g. Entity Tags) to inform the recipient when an update has occurred. The inefficiencies of polling are slightly less problematic than scheme 1 (there are likely to be fewer ReMs than embargoed resources), and this scheme hides the URLs of embargoed resources, which might be desirable. It would also be possible to modify the ReM in other ways than simply omitting resources – for example including a blank or redacted version of a chapter rather than omitting it completely (important for pagination if the parts are to be automatically reassembled). The primary downside to this scheme is additional complexity; the recipient needs a way of discovering that a ReM is partial, and when there are likely to be no more updates, probably by including additional data in the ReM itself.

Scheme 2: Send partial ReM, recipient polls ReM URL for updates

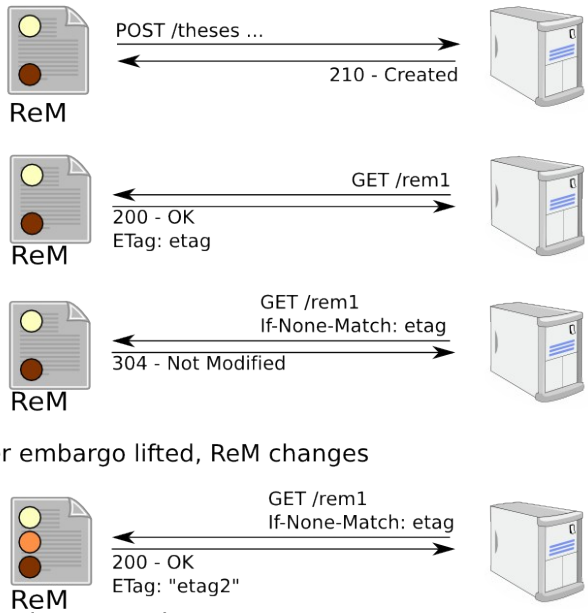


Illustration 16: Scheme 2

2.5.3 Scheme 3: Push Updated ReM

In the third scheme (Illustration 17: Scheme 3) the ReM is simply sent to the recipient whenever the embargo lifts on part of the thesis. This requires the sender and recipient to agree on an identifier in the ReM to identify the thesis, it makes sense to use the URI for the aggregation (the URI-a). This scheme is particularly suited for situations in which the recipient is to be the primary point of publication for the thesis, and has the additional advantages that (like scheme 2) it allows redacted versions of chapters to be substituted for completed ones, and that the sender is not required to act as a server, simply as a client.

Scheme 3: Send partial ReM followed by updated ReMs as embargo lifts

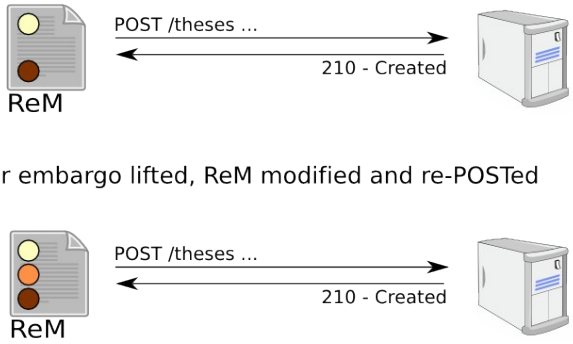


Illustration 17: Scheme 3

3 Technical summary

This section is a brief description of the technology used in this work. The main applications, ICE and The Fascinator are both evolving open systems, with websites that track ongoing development. Code and documentation is available for both.

ICE is an open source application written in the Python programming language, available for Windows, Mac and Linux platforms. I simplifies management of a set of files managed using the Subversion revision control system by providing the user with a web-view of content either on their local storage system or on a centralized server. ICE orchestrates content conversion from various file formats to web-ready formats using an extensible plugin system. It's original and main focus is on providing good quality HTML output for word processing documents using the OpenOffice.org application as part of the conversion system.

The Fascinator in a flexible repository component which provides a rich faceted index of content via the Apache Solr text indexer, with support for OAI-PMH ingest and dissemination. It is written in Java, with some interface and indexing plugins in Jython, to enable easy customization. The version used in the work reported here was tied to the FedoraCommons repository back-end but current versions offer a choice of back-end storage systems via an API, with a simple file-based storage layer currently available as an alternative to Fedora.

Evolved and tested versions of the SWORD + ORE work described here will be released as part of The Fascinator in 2010.

4 Conclusions

To summarize, innovations in the workflow/lifecycle of a thesis include:

1. Effective capture of metadata (technical and descriptive) as part of the authoring process rather as part of deposit process. In fact, the post-award deposit process has been replaced altogether in our proof of concept.
2. Showing how repository ingest can be made a by-product of an existing workflow, with data moving between systems based on the functional requirements of the stakeholders rather than a mandate to deposit data and papers. We contend that this direction whereby services are driven by the immediate motivations of the participants will be easier and quicker to bootstrap to a sustainable long-term business model than those driven by edict.
3. Working implementations of ORE – including code to both push content using SWORD and harvest it using the ATOM archive format which may be reused in other projects. This is achieved using metadata construction 'invisible' to the author, who is guided into creating good metadata and data through intuitive extensions to a familiar interface.
4. A proof-of-concept repository architecture for start-to-finish thesis management from authoring to dissemination, with an innovative approach to embargo management. This includes a nascent thesis repository built on Fedora-commons and The Fascinator (a Fedora front-end).

To summarize our work on workflow, ICE-TheOREM has followed existing academic workflows for authoring, examination, repository deposit. This work provides a proof-of-concept for true born digital web-eTheses. Embargo is handled by making sure that the requirements of the various stakeholders and parties are taken into account. Metadata about embargo is to be entered by the person best placed to know the requirement, the candidate, while we have recommended using an OpenId to identify the candidate so that the embargo can be managed even if they no longer have and institutional account.

The outcomes of the ICE-TheOREM project are summarized here:

- Open source code – available from USQ.
  - Extensions to the ICE content management system for OAI-ORE and Chemistry Markup Language.
  - ePrints and Fedora 3 modules for submitting HTML documents and packages via SWORD/OAI-ORE – now in use at USQ.
  - Extensions to the The Fascinator repository front-end for thesis embargo.
- A demonstration virtual machine with the project's outcomes on it for download (7GB) In VirtualBox VDI format (can be converted to use with VmWare)
- Openly available record of the development at the Cambridge Trac Wiki and at the Trac system at USQ.

The work reported here is a proof of principle for the ORE technology and a first step towards larger scale trials of repository-integrated thesis authoring workflows. A PhD thesis takes years to complete, so a true test of this infrastructure will involve a long term commitment. This commitment is being made at the Australian Digital Futures Institute – beginning in 2009 all the theses begin completed by institute staff and affiliates are housed in a system derived from the TheOREM work.

Further work starting now includes small scale trials with PhD candidates happening and conversion of recent theses into ICE at USQ. But much more work is required:

- Finish daily 'pull' of non-embargoed material from thesis repository to IR (work was started but not finished).
  - ATOM or ORE to show changes to embargo status
  - Dynamic building of Thesis PDF files omitting embargoed chapters.
  - SWORD + ORE as manifest with resources included.
- Work on managing thesis examination process with possible online submission of reports (at USQ OJS has been used for this in the Maths and computing department).
- OAI-ORE + SWORD gives us part of the puzzle but agreed content models for theses, journals etc are still needed.
- Investment is required in the Graduate Studies repository and its workflows.
- Solutions are needed for allowing repositories to *optionally* provide added-value services (like 3d molecules) while degrading gracefully, with only declarative markup embedded in repository items ('this is chemistry')

rather than code to invoke particular viewing software.

5 References

Downing, J., Allinson, J. & et al, 2009. SWORD AtomPub Profile. Available at: http://swordapp.org/sword/specifications [Accessed September 15, 2009].

Jacobs, N., 2008. Departmental Thesis Management System development using the Integrated Content Environment (TheOREM-ICE). Available at: http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/theorem-ice.aspx [Accessed July 14, 2008].

Monus, L. et al., 2008. Zero Click Ingest. Available at: http://pubs.or08.ecs.soton.ac.uk/119/ [Accessed May 20, 2008].

Murray-Rust, P., 2007. The Power of the Electronic Scientific Thesis. *10th International Symposium on Electronic Theses and Dissertations*. Available at: http://epc.ub.uu.se/ETD2007/sessions/keynote-2.html?keepThis=true&TB_iframe=true&height=480&width=640 [Accessed September 8, 2009].

Murray-Rust, P. & Rzepa, H.S., 2004. The Next Big Thing: From Hypermedia to Datuments. *Journal of Digital Information*, 5(1), 248. Available at: http://journals.tdl.org/jodi/article/viewArticle/130/128

Neylon, C., 2008. Science in the open » A personal view of Open Science - Part IV - Policies and standards. Available at: http://blog.openwetware.org/scienceintheopen/2008/10/26/a-personal-view-of-open-science-part-iv-policies-and-standards/ [Accessed February 5, 2009].

Recordon, D. & Reed, D., 2006. OpenID 2.0: a platform for user-centric identity management. *Proceedings of the second ACM workshop on Digital identity management*, 11-16. Available at: http://portal.acm.org/citation.cfm?id=1179532

Sefton, P., 2006. The integrated content environment. In *AUSWEB 2006*. Noosa: Southern Cross University. Available at: http://eprints.usq.edu.au/archive/00000697/01/Sefton_ICE-ausweb06-paper-revised-3.pdf .

Sefton, P., Downing, J. & Day, N., 2009. ICE-theorem - end to end semantically aware eResearch infrastructure for theses. *University of Southern*

Queensland. Available at: http://eprints.usq.edu.au/5248/1/ice-theorem-paper-OR09.htm [Accessed August 24, 2009].

Sefton, P. & Lucido, O., 2009. The Fascinator: a lightweight, modular contribution to the Fedora-commons world. In Atlanta, Georgia. Available at: http://eprints.usq.edu.au/5259/ .