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

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

Jim Downing
University of Cambridge
ojd20@cam.ac.uk

Nick Day
University of Cambridge
ned24@cam.ac.uk

Oliver Lucido
University of Southern Queensland
lucido@usq.edu.au

2009-05-27

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 adding 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 publication 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.

One the technological side, ICE-TheOREM implemented the Object Reuse and Exchange (ORE) protocol to integrate 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, which 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.

Introduction

Acknowledgements & Credits

ICE-Theorem was a joint project between the University of Cambridge (UC) and the University of Southern Queensland (USQ) funded by the JISC (Jacobs 2008)

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, [TODO: JD]

Project motivations

[TODO: JD]

In this paper we follow the workflow of writing and supervision, examination and deposit of a thesis showing where the ICE-TheOREM proj ect[1] 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.

Project outcomes

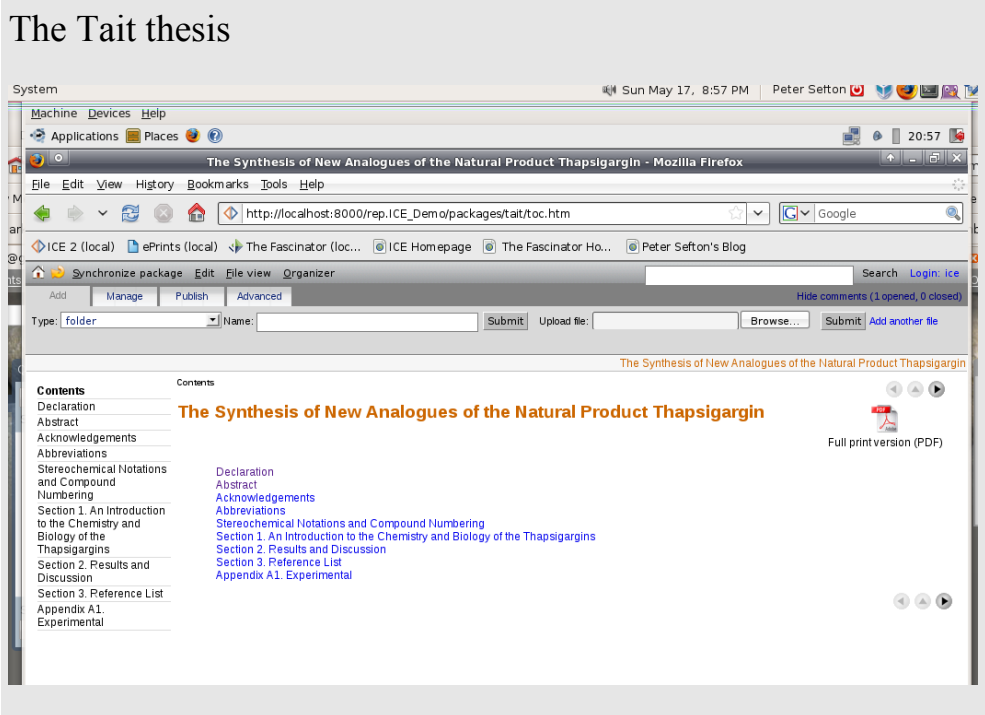
- Open source code – available from USQ.
  - Extensions to the ICE content management system for OAI-ORE and Chemistry Markup Language.
  - ePrints and Fedora 3 modules for submitting HTML documents and packages via SWORD/OAI-ORE – now in use at USQ.
  - Extensions to the The Fascinator repository front-end 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. xxxx

The TheOREM project aimed to exercise the OAI-ORE protocol (IONSREPORT 2008) in the context of chemical theses – with content . The ICE (Integrated Content

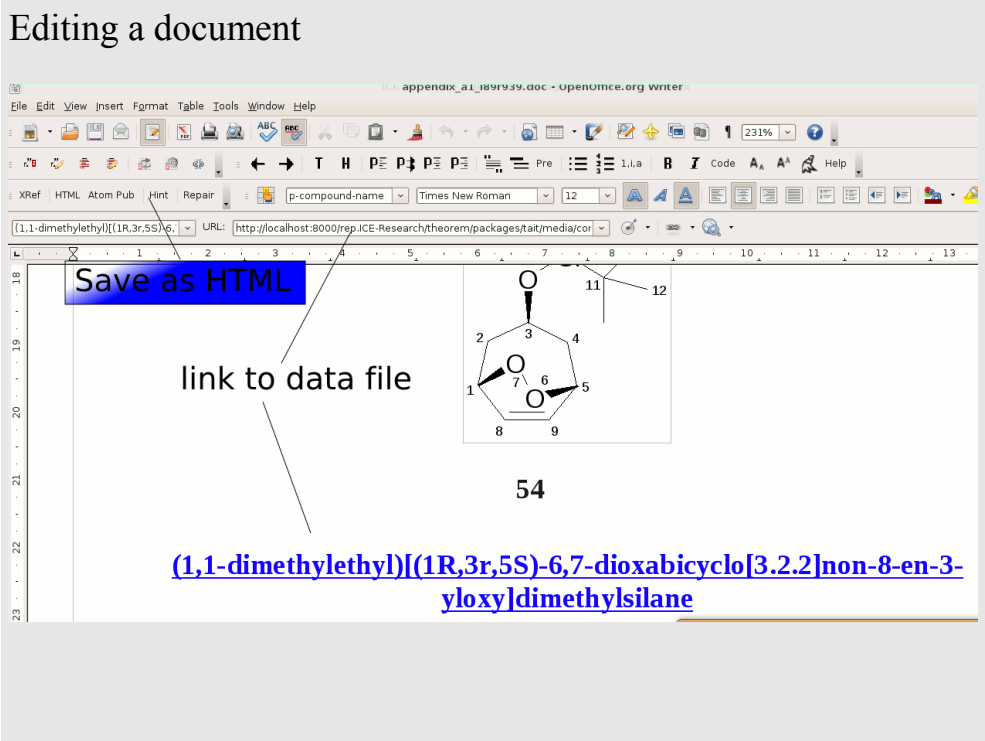
1 This is a footnote

Environment) 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 [TODO: ref]. 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) that is they are hypertext aggregations of document and data, which are both human and machine-readable.

The centerpiece of the ICE-TheOREM project has been a thesis by Malcolm Tait. The thesis is show here in the ICE system, running in the virtual machine we created for the project:

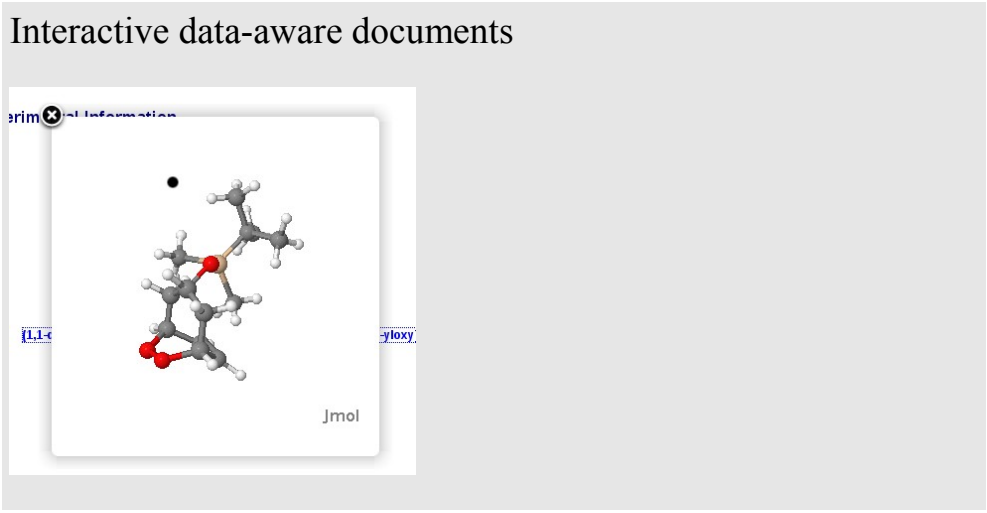


Each of the documents that make up the thesis is a word processing file. In this case Microsoft Word (.doc) format, but OpenOffice.org (.odt) files are also supported.

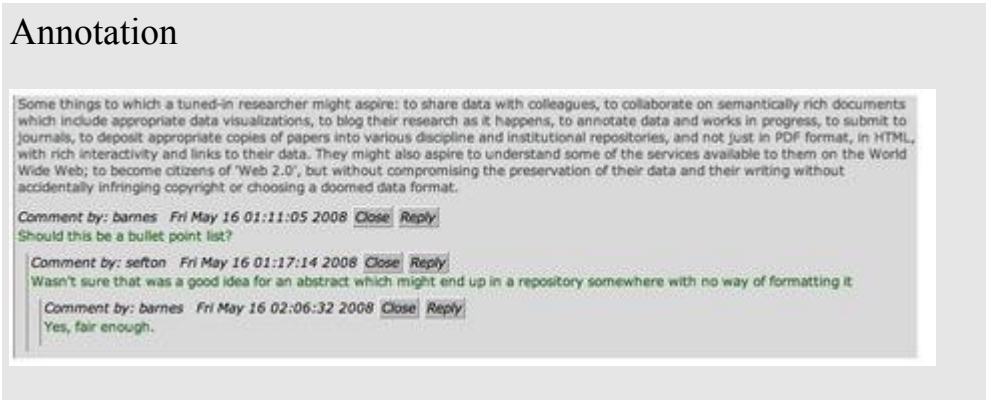


The key features of an ICE document are highlighted in the above screenshot. It uses styles to convey structural information about a document, the author applies styles using a toolbar, and the document can be converted to HTML format or sent to a website (usually a weblog) via the Atom Publishing protocol. 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.

Because 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, the ICE application embeds a 3d rendition of the molecule of the page.

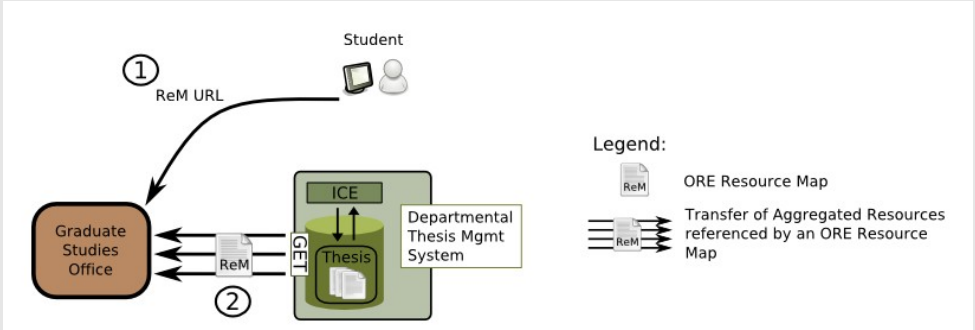


The ICE system allows for stand-off annotation of documents in a way that is similar to the ComentPress system. Supervisor(s) and peers to comment on a document without changing it.



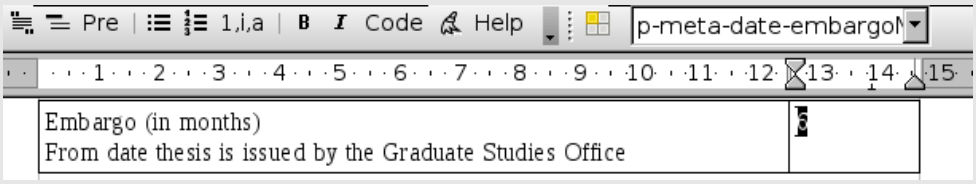
When the document is ready for submission for examination, the ICE-TheOREM model proposes a repository which belongs to the graduate studies office, so the thesis needs to be deposited in that repository. This could be accomplished by a 'pull' process where the repository watches the ICE system and fetches theses with a certain flag set, such as ready for examination, as described in the competition entry for Open Repositories 2008 *Zero Click Ingest*, but in ICE-TheOREM we have used a push system, where the candidate uses the SWORD function to send the thesis to a thesis repository.

Initial thesis submission - schematic



One of the major contributions of ICE-TheOREM is a model for granular thesis embargo. Our demonstration image contains one document which is to be embargoed. In this case it is the acknowledgements section which will embargoed, it is more likely that more substantial parts of a thesis will be embargoed for reasons to do with commerical exploitation or privacy of subjects, but 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.

Embargo metadata is encoded in a style:

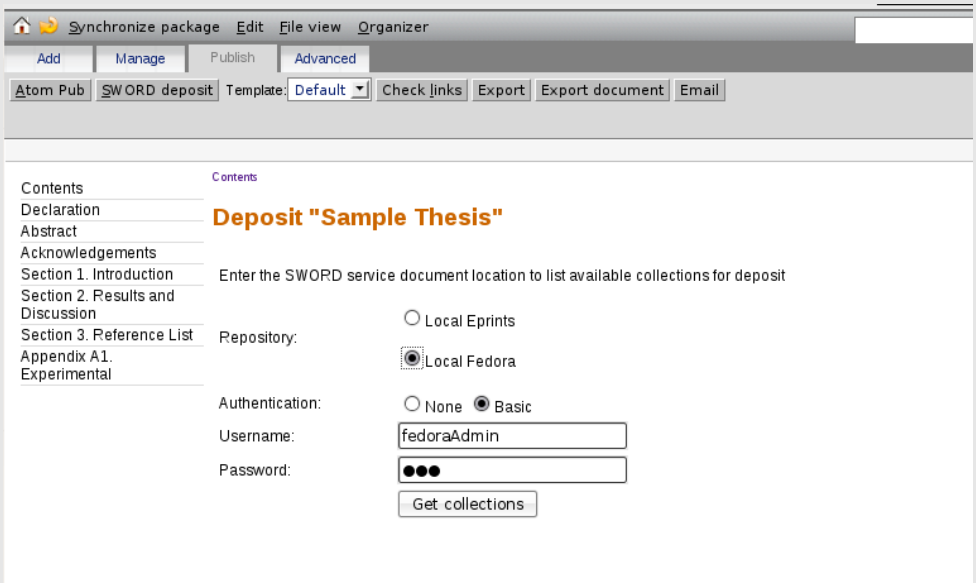


And ICE can extract the metadata:

```
<oai_dc:dc>
<dc:title>Acknowledgements</dc:title>
<dc:relation>date-embargoMonths::6</dc:relation>
</oai_dc:dc>
```

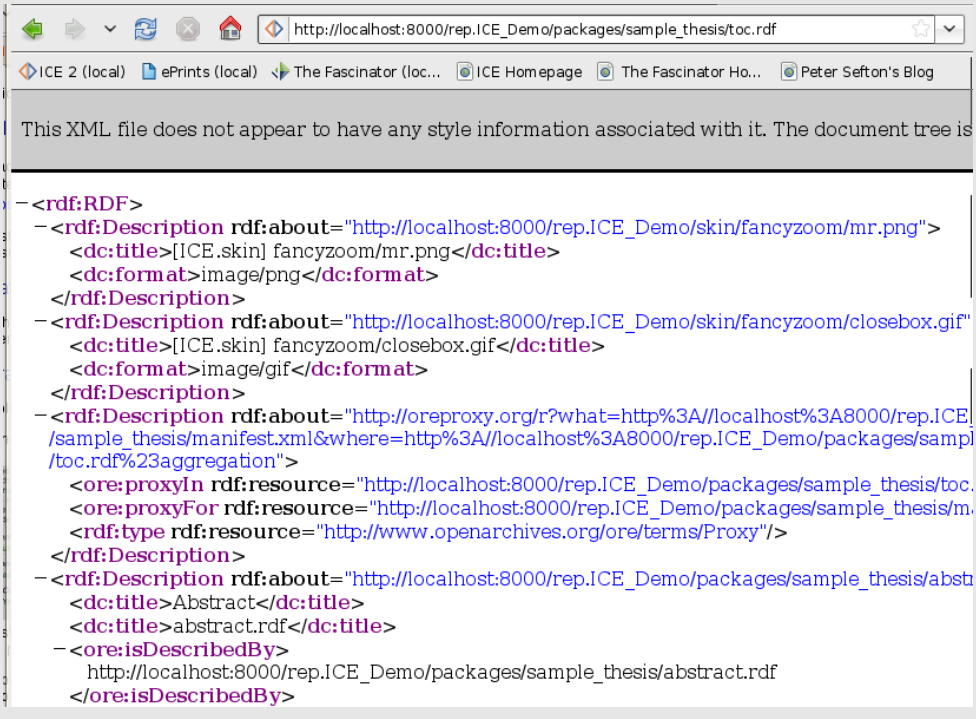
When the thesis is sent to the thesis repository via SWORD, then the metadata is sent with it. We propose that the graduate studies office get the student to submit an OpenId – allowing them to administer embargoes using the OpenId when their institutional login may have expired.

SWORD deposit



The SWORD deposit contains an OAI-ORE payload.

SWORD deposit uses an ORE Resource Map



This XML is expressing the structure of the thesis.

ORE for a thesis

[TODO: Jim – can you draw an ORE RM for a thesis with HTML parts, PDF parts and CML linked into some of the parts?]

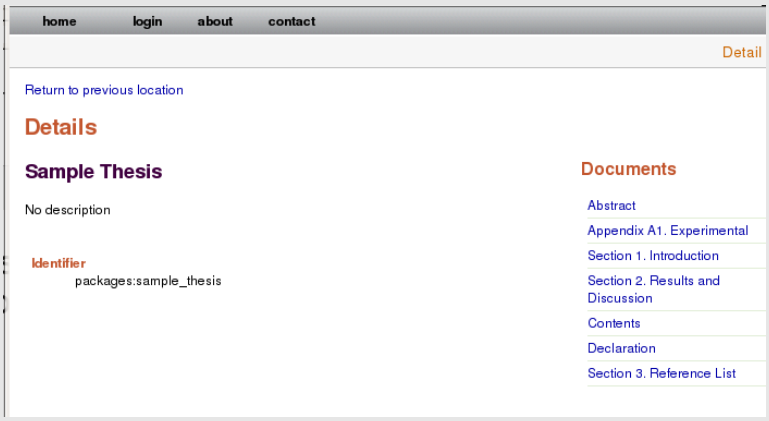
Importance of ORE

- 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.)

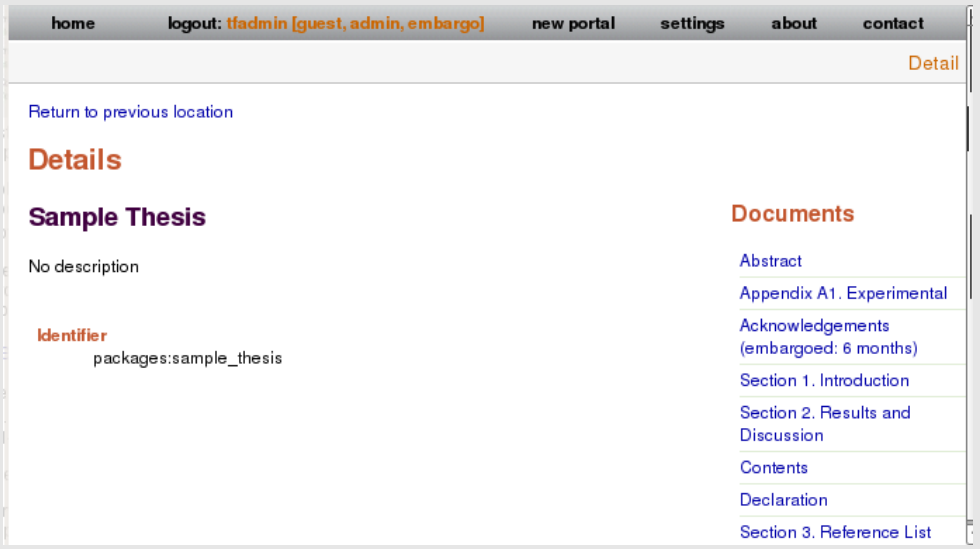
The thesis repository is currently only a mock-up, using The Fascinator to serve theses from a Fedora repository.

Default view of thesis – note embargoed chapter is not seen



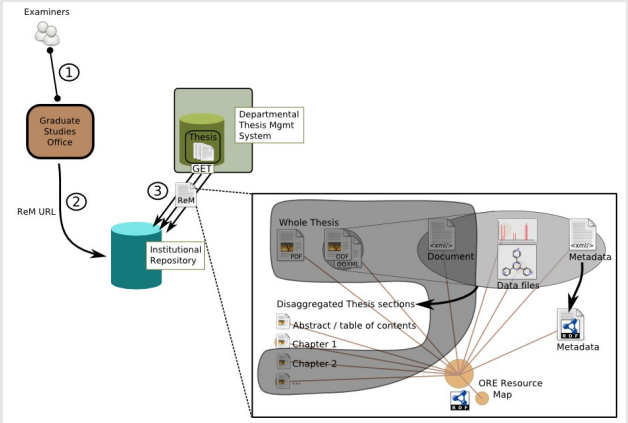
Whereas if an administrator is logged in then the acknowledgements are visible.

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.

The final stage – automated IR deposit



More work needed on thesis repository

- Finish daily 'pull' of non-embargoed material from thesis repository to IR (work was started but not finished)
- 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).

To summarize, innovations in the workflow/lifecyle 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 based on OpenID. This includes a nascent thesis repository built on Fedora-commons and The Fascinator (a Fedora front-end).

Workflow summary

- ICE-TheOREM has followed existing academic workflows
  - Authoring
  - Examination
  - Repository deposit
- Provides a proof-of-concept for true born digital web-eThese

Conclusion: Further work required

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 early 2009 all the theses begin completed by institute staff and affiliates will be housed in a system derived from the TheOREM work (Observatory PASCAL http://www.obs-pascal.com/).

Further work starting now

- Small scale trials with PhD candidates happening at USQ now
- Conversion of recent theses into ICE at USQ now underway

More work needed

[TODO: PS]

References

IONSREPORT, S., 2008. OAI-ORE specifications. *Scholarly Communications Report*, 12(1), 5-5.

Jacobs, N., 2008. Small-Scale OAI Object Re-Use and Exchange Experiments. Available at: http://www.jisc.ac.uk/whatwedo/programmes/digitalrepositories2007/oaioredemonstrators.aspx [Accessed February 5, 2009].

Murray-Rust, P. & Rzepa, H.S., 2004. The Next Big Thing: From Hypermedia to Datuments. *Journal of Digital Information*, 5(1), 248.

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].