

# 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

## 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 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.

On 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

This document is the basis for a presentation for OR09, using [ICE](#) to embed presentation material in a document. This will be posted to USQ ePrints and later developed into a full paper for the conference proceedings.

## 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.

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 main man, with support from Joe Townsend.

## Motivations: Put ORE through its paces

- Is it applicable? Is it useful?
- What are the different ways of using ORE?
- How do SWORD and ORE combine?

## SWORD + ORE options

- ORE as a manifest (with resource included)
- ORE as a shopping list (target orders all the resources)  
this is what we implemented
- ORE as a recipe (tells you how, but you don't have to get the resources)

## Motivations: Design a thesis workflow based around web architecture

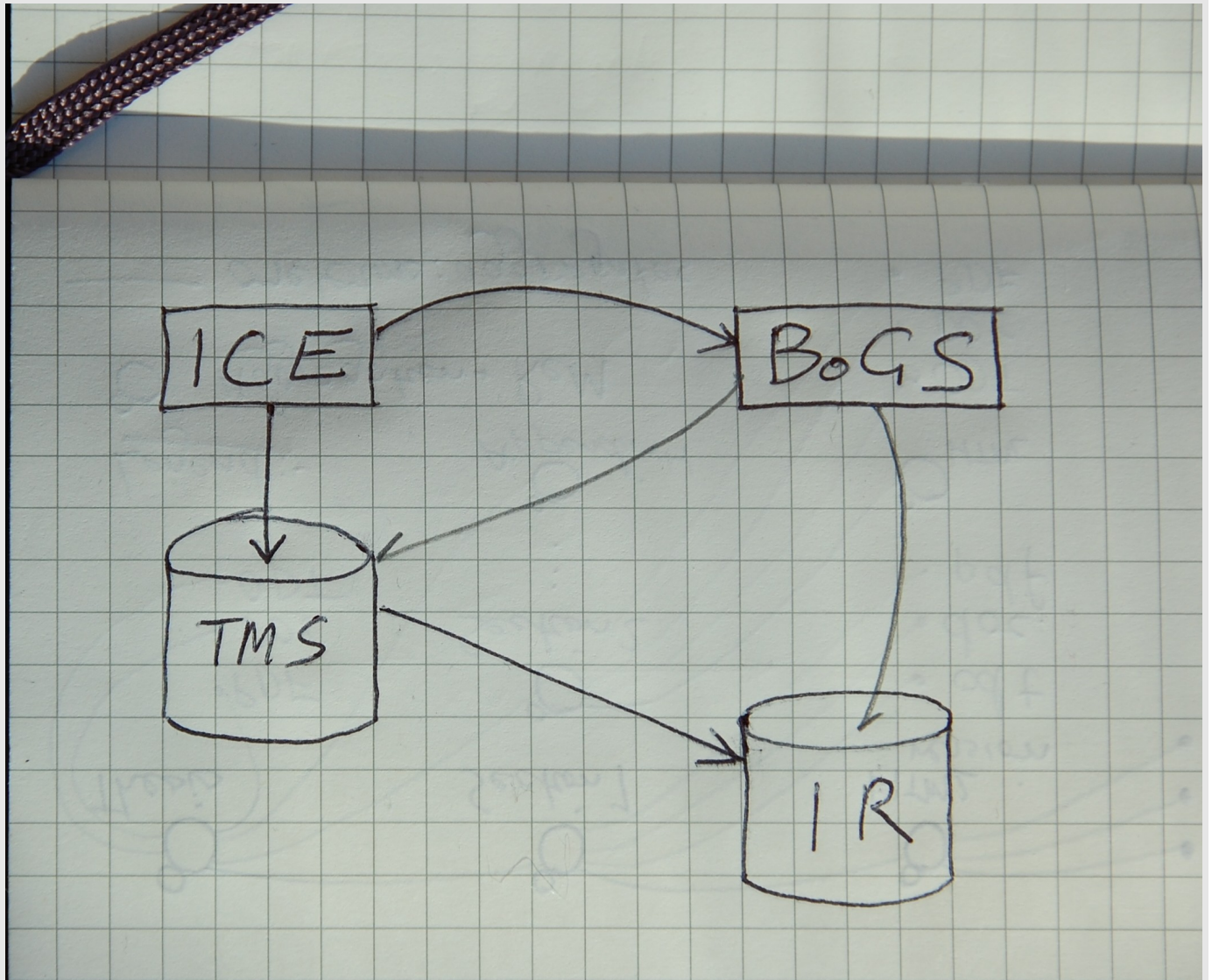
- Interoperability for easier integration.
- The distributed IR “It's a state of mind”

## Motivations: Work out whether disaggregation of theses could promote open access

- Promoting embargo of sensitive chapters might accelerate publication of the remainder of the thesis.
- 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.

In this paper we follow the workflow of writing and supervision, 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.

# Overall thesis workflow with thesis repository and Board of Graduate Studies (BoGS)



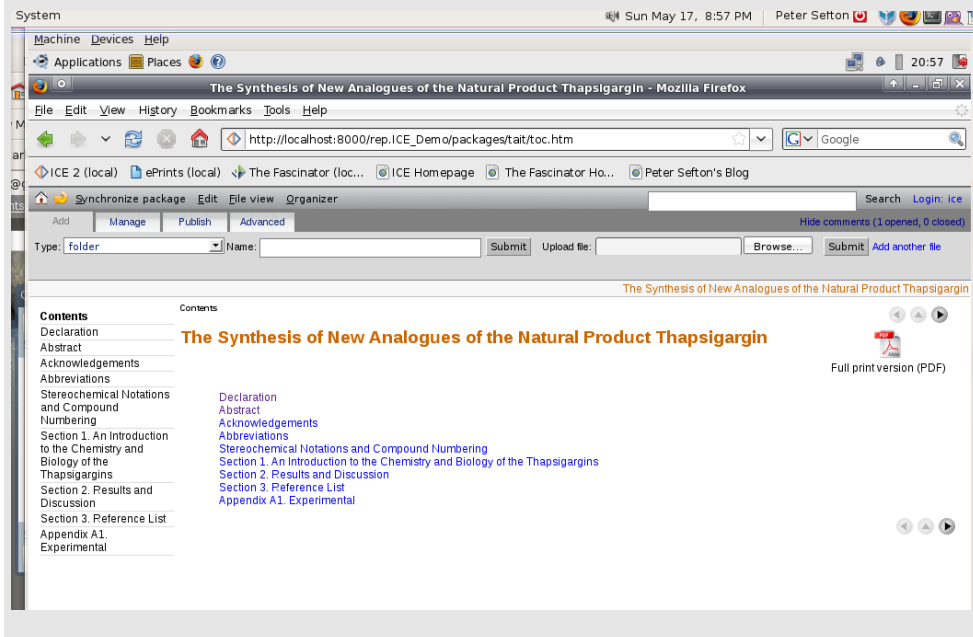
## Project outcomes

- 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](#).

The TheOREM project aimed to exercise the OAI-ORE protocol (IONSREPORT 2008) in the context of chemical theses – with contents The [ICE](#) (Integrated Content Environment) (P 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 . 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 shown here in the ICE system, running in the virtual machine we created for the project. 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.

## The Tait thesis



## Editing a document

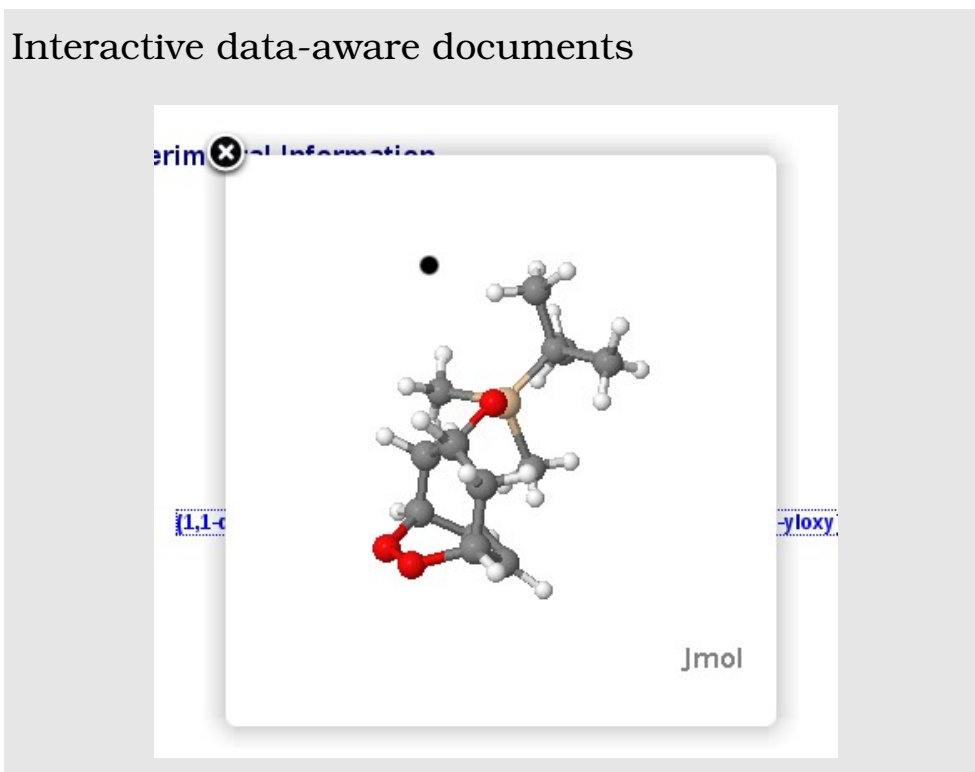
A screenshot of the OpenOffice.org Writer application editing a document. The document title is "ICE appendix\_a1\_189f939.doc". The interface shows a menu bar, a toolbar, and a text area. The text area contains a chemical structure diagram of a bicyclic compound with numbered atoms (1-12) and the number "54" below it. The chemical structure is a bicyclic system with two oxygen atoms and a dimethylsilane group. The text below the structure is (1,1-dimethylethyl)[(1R,3r,5S)-6,7-dioxabicyclo[3.2.2]non-8-en-3-yloxy]dimethylsilane. Annotations with arrows point to various features: "Save as HTML" points to the "HTML" button in the top toolbar; "link to data file" points to the "Atom Pub" button; "Formatting toolbar which applies styles" points to the main formatting toolbar; and "URL: http://localhost:8000/rep.ICE-Research/theorem/packages/taut/media/cor" is shown in the address bar.

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 – 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.

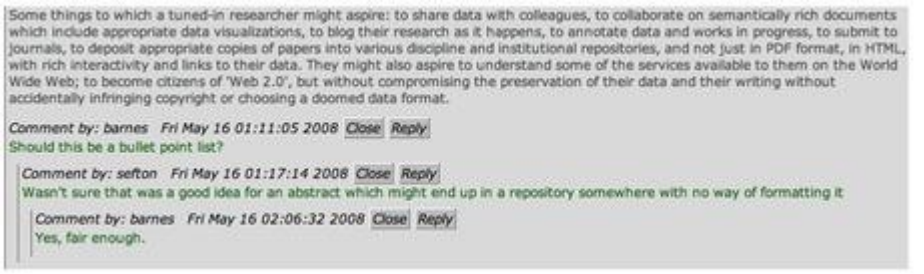
In this case, the image of a molecule, (1,1-dimethylethyl)[(1R,3r,5S)-6,7-dioxabicyclo[3.2.2]non-8-en-3-yloxy]dimethylsilane is linked to a Chemical Markup Language file describing it so the ICE application embeds a 3d rendition of the molecule of the page in its web-renderition. This data accompanies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

## Interactive data-aware documents



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

## Annotation

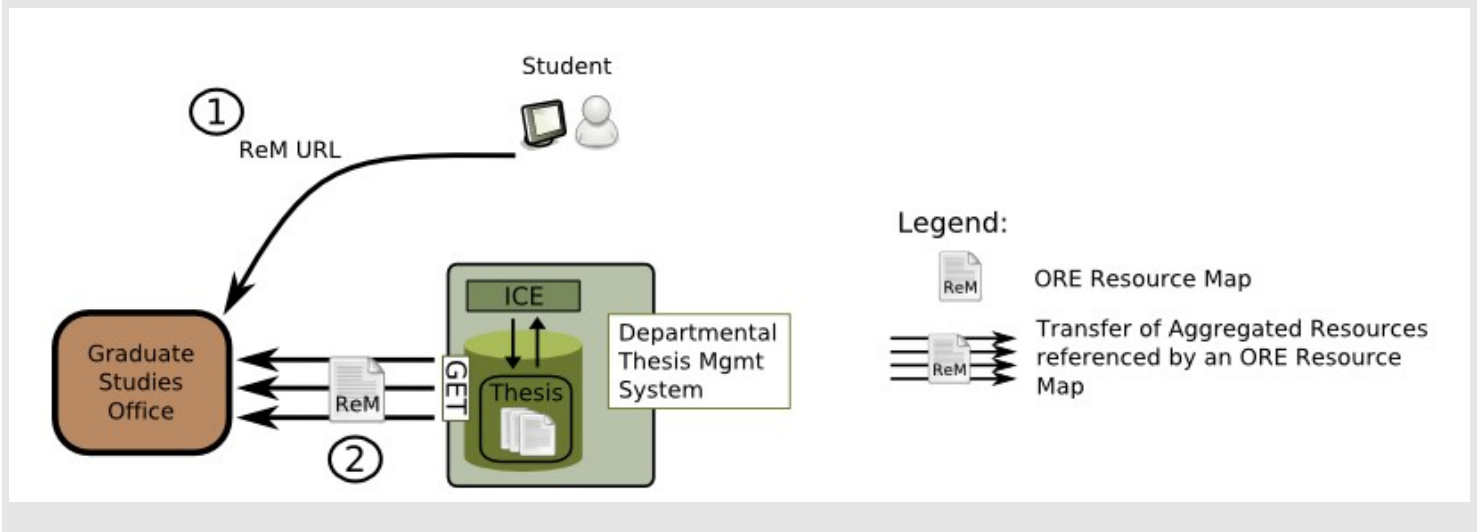


When the document is ready for submission for examination, the ICE-TheOREM model proposes a

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 *read\_for\_examination*, as described and prototyped in the competition entry for Open Repositories 2008 *Zero Click Ingest (Monus et al. 2008)*, 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. 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.

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.

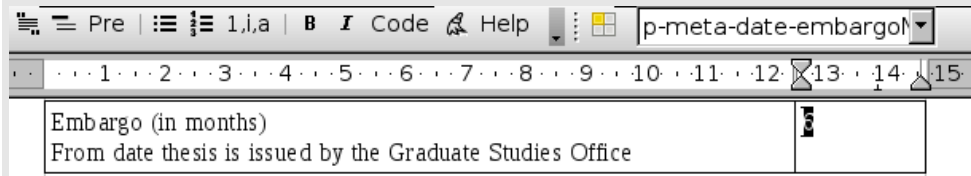
## Initial thesis submission - schematic



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. In this demonstration we have one document which is to be embargoed. In this case it is the acknowledgements section. While it is more likely that more substantial parts of a thesis will be embargoed for reasons to do with commercial exploitation or privacy of subjects, 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:

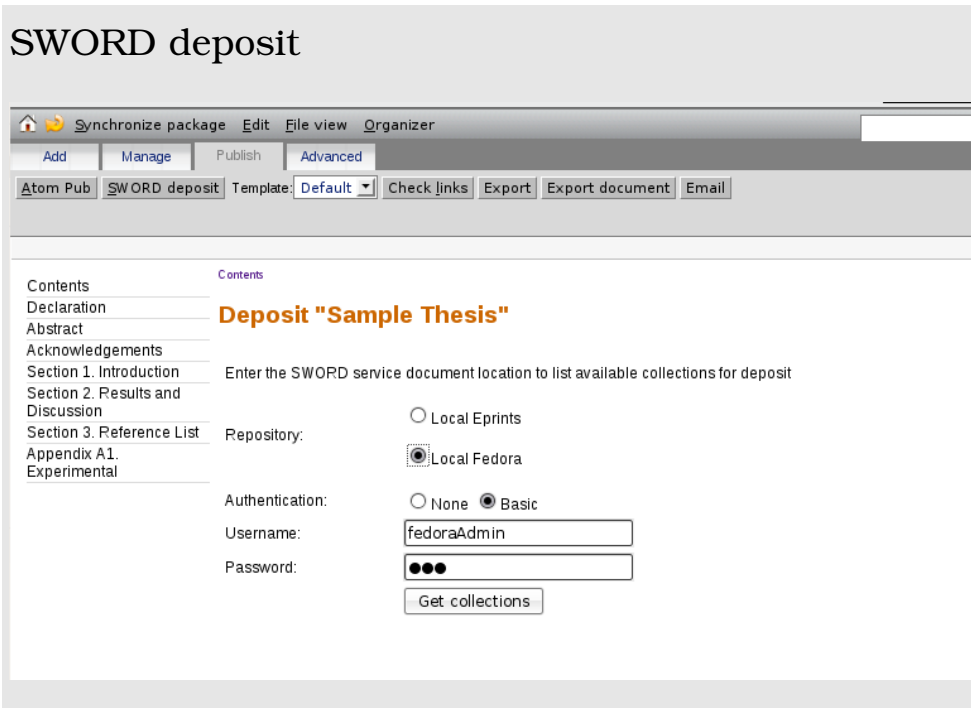


And ICE can extract the metadata:

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

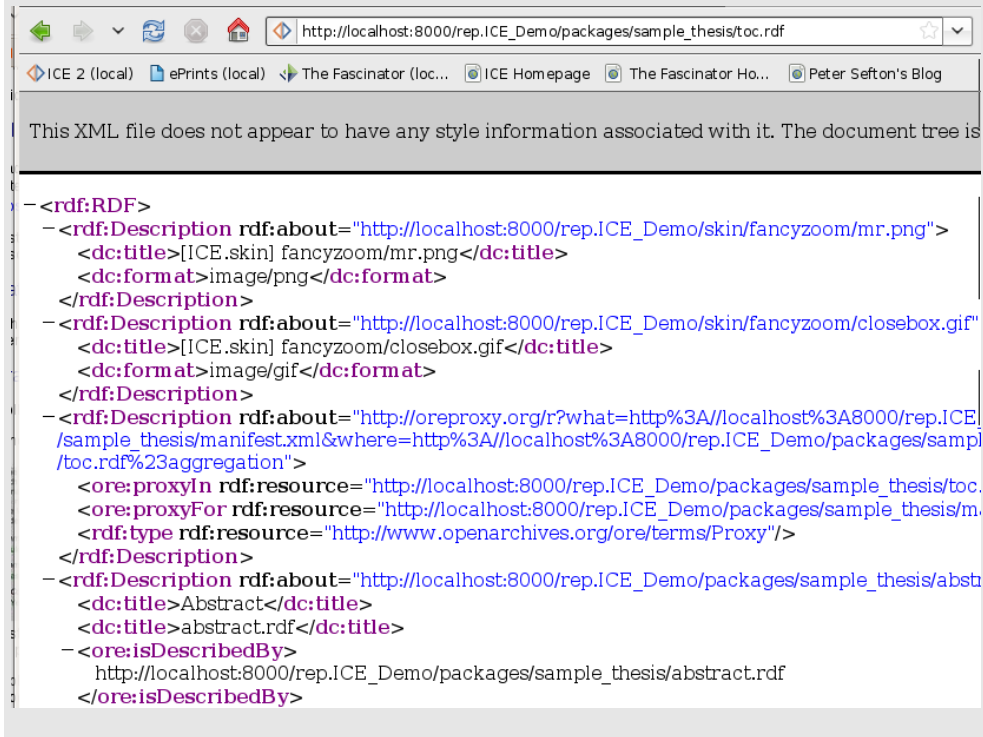
The initial demonstration encodes embargo information using a style, using a technique developed in the ICE and ICE-TheOREM projects.

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 an OpenId – allowing them to administer embargoes using the OpenId when their institutional login may have expired, 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.



The SWORD deposit contains an OAI-ORE payload.

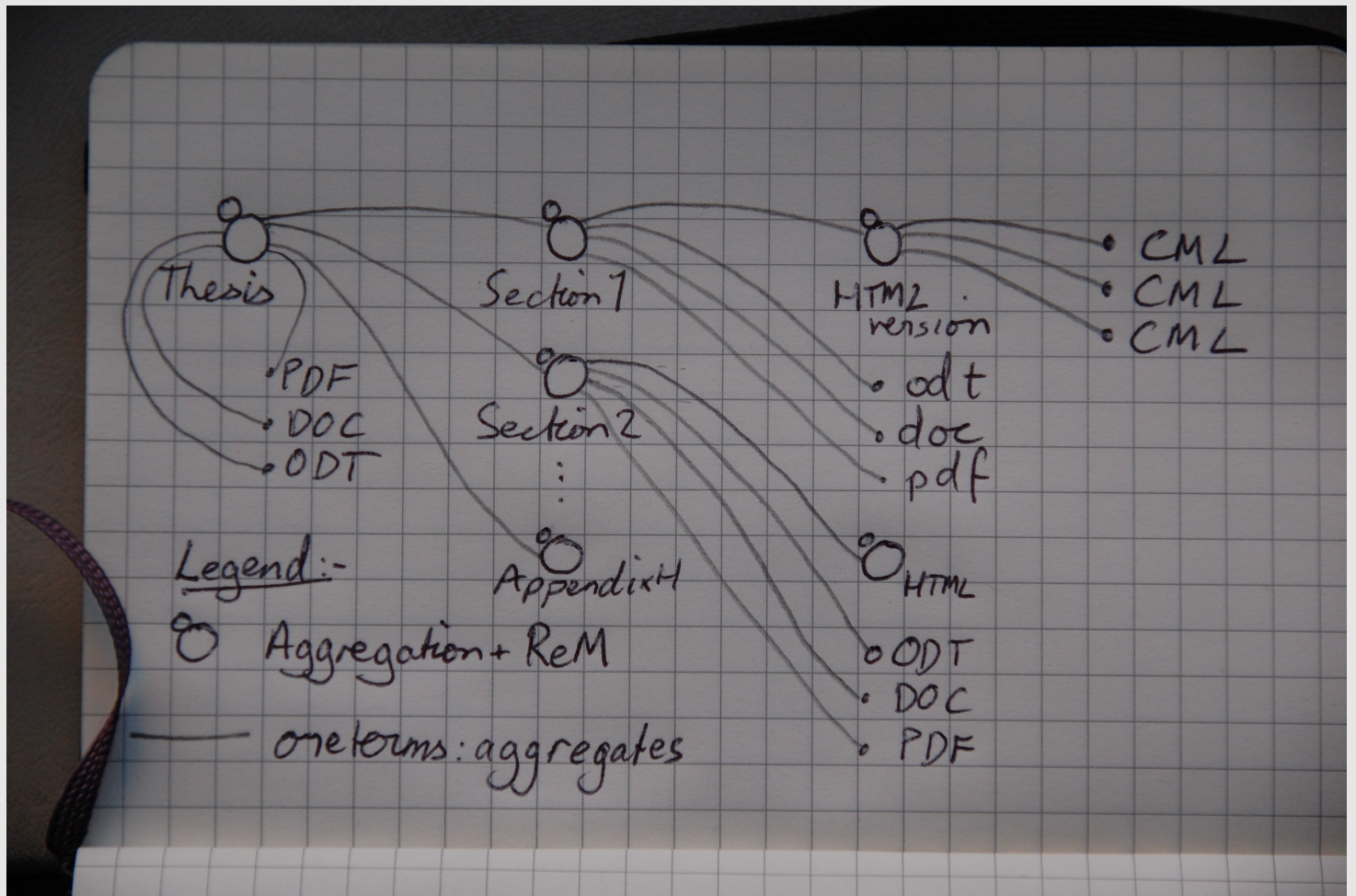
## SWORD deposit uses an ORE Resource Map



```
-<rdf:RDF>
- <rdf:Description rdf:about="http://localhost:8000/rep.ICE_Demo/skin/fancyzoom/mr.png">
  <dc:title>[ICE.skin] fancyzoom/mr.png</dc:title>
  <dc:format>image/png</dc:format>
</rdf:Description>
- <rdf:Description rdf:about="http://localhost:8000/rep.ICE_Demo/skin/fancyzoom/closebox.gif">
  <dc:title>[ICE.skin] fancyzoom/closebox.gif</dc:title>
  <dc:format>image/gif</dc:format>
</rdf:Description>
- <rdf:Description rdf:about="http://oreproxy.org/r?what=http%3A//localhost%3A8000/rep.ICE_Demo/sample_thesis/manifest.xml&where=http%3A//localhost%3A8000/rep.ICE_Demo/packages/sample_thesis/toc.rdf%23aggregation">
  <ore:proxyIn rdf:resource="http://localhost:8000/rep.ICE_Demo/packages/sample_thesis/toc.rdf">
  <ore:proxyFor rdf:resource="http://localhost:8000/rep.ICE_Demo/packages/sample_thesis/manifest.xml">
  <rdf:type rdf:resource="http://www.openarchives.org/ore/terms/Proxy"/>
</rdf:Description>
- <rdf:Description rdf:about="http://localhost:8000/rep.ICE_Demo/packages/sample_thesis/abstract.rdf">
  <dc:title>Abstract</dc:title>
  <dc:title>abstract.rdf</dc:title>
  <ore:isDescribedBy>
    http://localhost:8000/rep.ICE_Demo/packages/sample_thesis/abstract.rdf
  </ore:isDescribedBy>
```

This XML is expressing the structure of the thesis.

## ORE for a thesis



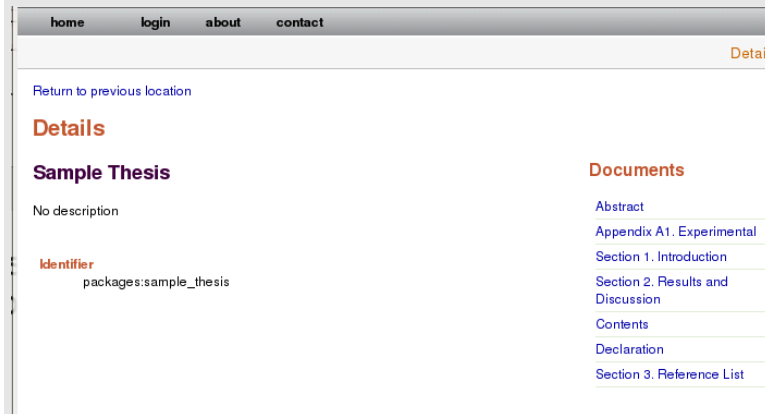
### Importance of ORE

- 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.)

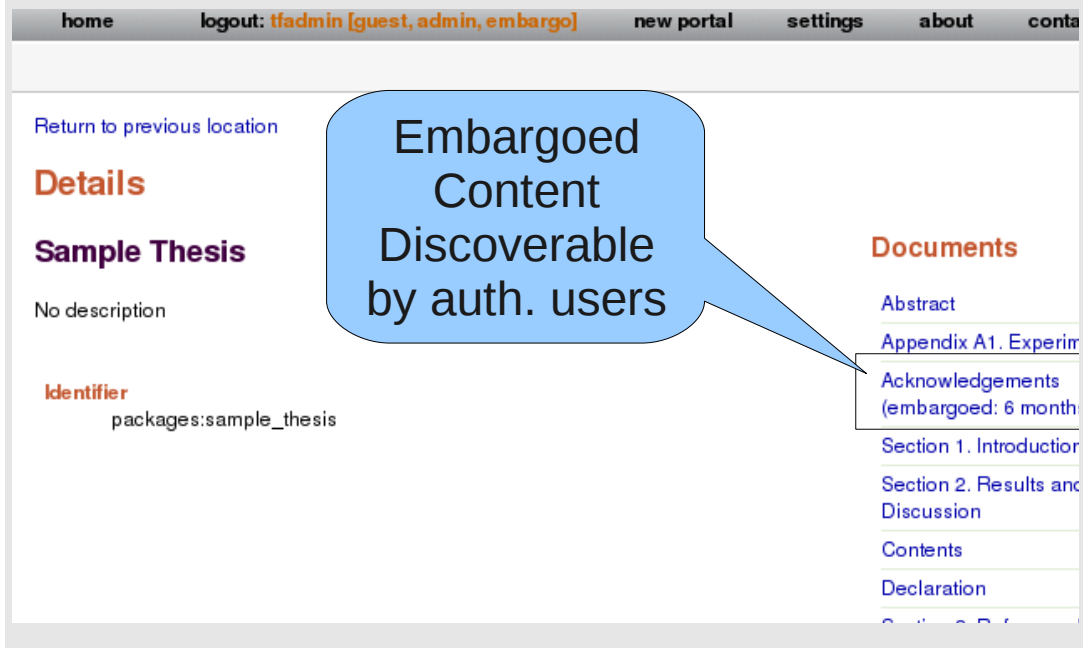
The thesis repository is currently only a mock-up, using [The Fascinator](#) to serve theses from a Fedora 3 repository.

## Default view of thesis - no acknowledgements



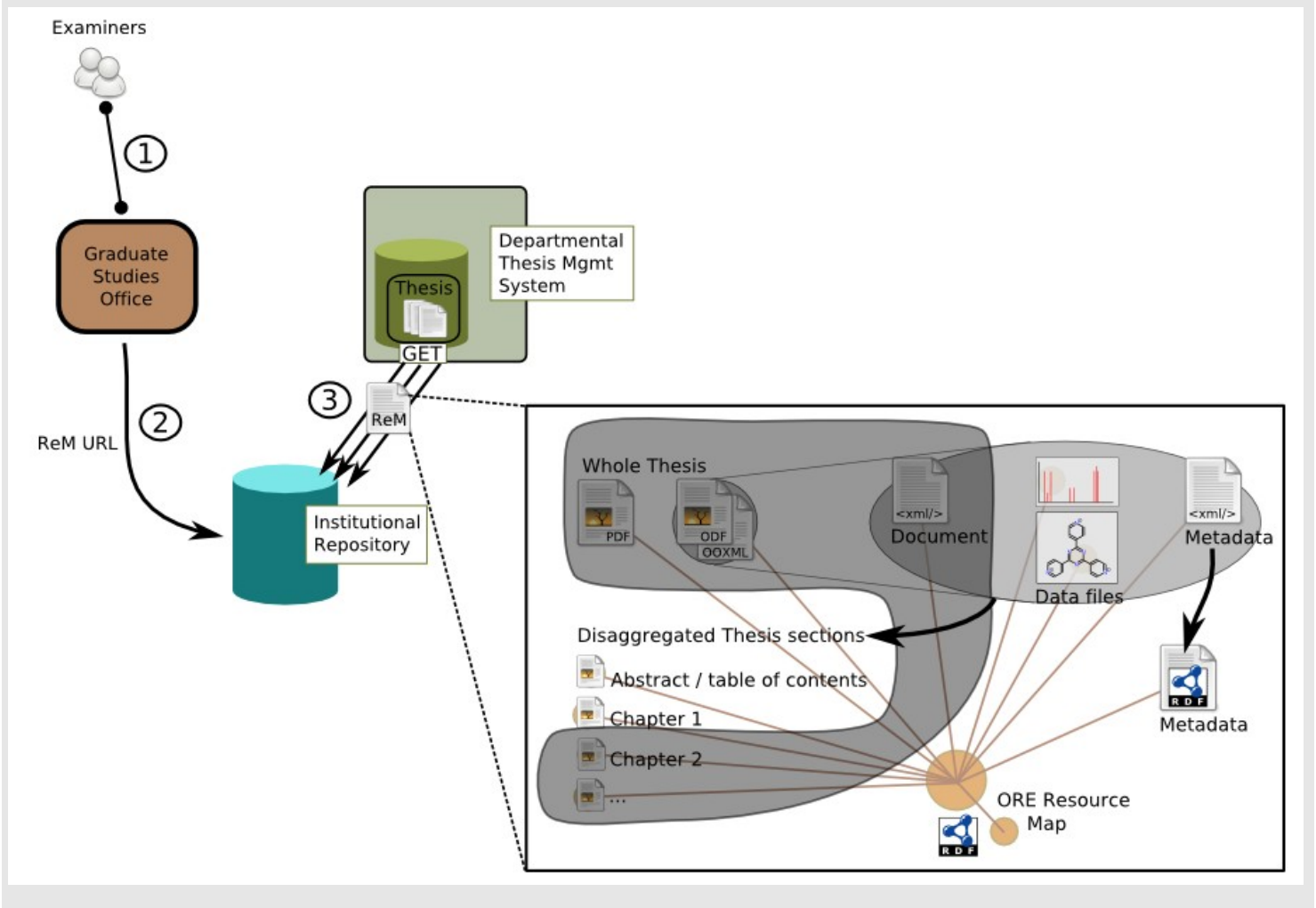
Whereas if an administrator is logged in then the acknowledgements are visible.

## 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



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 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
  - Embargo administered by the student's OpenId
- 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.

## 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 on thesis repository

- 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).

## Help wanted!

- 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 needed for allowing repositories to *optionally* provide added-value services (like 3d molecules) while degrading gracefully.

## References

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

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. & Rzepa, H.S., 2004. The Next Big Thing: From Hypermedia to Datuments. *Journal of Digital Information*, 5(1), 248. Available at: <http://jodi.tamu.edu/Articles/v05/i01/Murray-Rust/?printable=1>.

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].

Sefton, P., 2006. The Integrated Content Environment for Research and Scholarship. *ICE Website*. Available at: [http://ice.usq.edu.au/introduction/ice\\_rs.htm](http://ice.usq.edu.au/introduction/ice_rs.htm) [Accessed April 30, 2007].