What the OAI-ORE protocol can do for you

Peter Sefton

University of Southern Queensland

sefton@usq.edu.au

A presentation for the ARROW Repository day 2008-12-14

Abstract: Open Archives Initiative Object Reuse and Exchange (OAI-ORE), is an important new protocol for representing compound objects, or aggregations, in a web environment. The system is generating a lot of development activity in the repository community some of which will be reported in this presentation.

One of the main contributions of ORE will be a way to describe an item that is made up of several parts. The classic example is an HTML document and its images – until the advent of OAI-ORE there has been no standardized way to draw a line around such an aggregation. What may seem obvious to a reader has not been obvious to machines for which the document and its images are treated as equivalent in status. Likewise ORE will help to define what is an item in a repository in a way that can help to make items portable between systems. It will also allow systems to exchange objects that are made up of multiple parts, such as a thesis with multiple chapters and data files.

The presentation will include some ORE demonstrations and discussion – showing some USQ work on how documents from a content management system can be automatically ingested into repository systems (we will demo with ICE, ePrints and The Fascinator), and how items can be migrated from one repository to another. There will also be some more speculative discussion of future possibilities and some examples of other work.

OAI is a conceptually complex system with its own very tightly defined set of terms and some very refined and nuanced design. It is likely that for the most part, developers will work with OAI-ORE libraries to get things done, and for end users and repositarians the system will be completely transparent in much the same way most of us never need to look *inside* an OAI-PMH feed.

1 Introduction – Standards

Before I go on to talk about ORE – that's Object Reuse and Exchange – I thought I'd talk a little bit about standards in general and then about some experiences with standards in the Australian repository domain, particularly in the ARROW community.

I used to work at Standards Australia. They liked to say there “The good thing about standards is that there are so many of them.” It was a laugh a minute, at Standards at least when it was in Homebush in the good old days, before they moved it to the CBD and floated it on the stock exchange.

Me, I always wanted a little sign in the office that said “We have standards to maintain”.

Eventually I got sick of wearing a tie and moved to Queensland.

So what's the point of a Standard?

In the software world one thing people worry about is, can I replace this bit of software with a different one later? Can I shop around for a database? Even when we're dealing with open standards and open software he answer is often, *well sort of*.

Standard components

You might be able to drop in / screw in components such as:

1. A house brick.
2. A tap.
3. A database (given the right drivers). Eg Fedora works with several different databases.
4. A screw (given the right drivers)
5. An authentication system like LDAP.

It's not just about components, though. The big thing is interoperability. Can my systems interoperate with yours now? And will my digital assets interoperate with **my own** future systems?

The thing that's important about those standard components is their interfaces. That is, how they fit together.

And how they behave if you move them from one site to another.

I make the point about interop because I want to come back to it later. You need to think very hard about whether it is worth using a standard to do something that you can't later reuse.

2 Two examples of standards

Lets look at two standards that have been used and/or promoted in the Repository space before going on to look at OAI-ORE; OAI-PMH and XACML.

2.1 OAI-PMH

The Open Access Initiative Protocol for Metadata Harvesting is a must-have standard for repositories. It's used for disseminating repository content to registries and indexes that aggregate content.

OAI-PMH

- Mostly works.
- A standard for moving metadata from one place to another.

But:

- **It's just the messenger.**

  Not all the messages that people send over it are coherent.
- **There are ugly bits.**

  Particularly its use of non-standard Dublin Core identifiers.

OAI-PMH is, basically A Good Thing. That's not to say that it's painless, though.

The OAI-PMH standard actually clashes with the Dublin Core standard in at least one place. Neil Godfrey nails the issue in a blog post.

But worse than that in practical terms harvesting is a mess. While the interchange protocol (PMH) more or less works the stuff that people interchange is very far from being standardized. At the National Library of Australia's ARROW discovery service they have a normalization process built into their harvester that presents a coherent view. This is not the result of standardization, it's the result of Alison Dellit and the NLA team's hard work in writing rules that say thing like:

```
For the USQ repository:
  If type is_equal_to   ADT_Thesis":
     set type to "australasian digital thesis
```

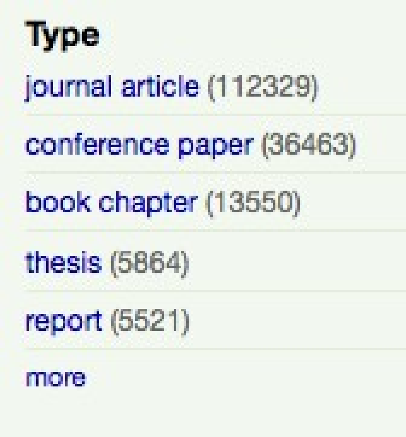
These rules normalize the chaos that is Australian Institutional Repositories. This is certainly made easier by a pretty-good level of standardization in some areas. At least people put the resource type in the dc:type element!

Don't shoot the messenger (PMH)

The Australian University Repository Census (AUIRC – pronounced OIK) uses OAI-PMH to harvest items from as many Australian Universities as it can.

Compare!

ARROW search



Type

journal article (112329)

conference paper (36463)

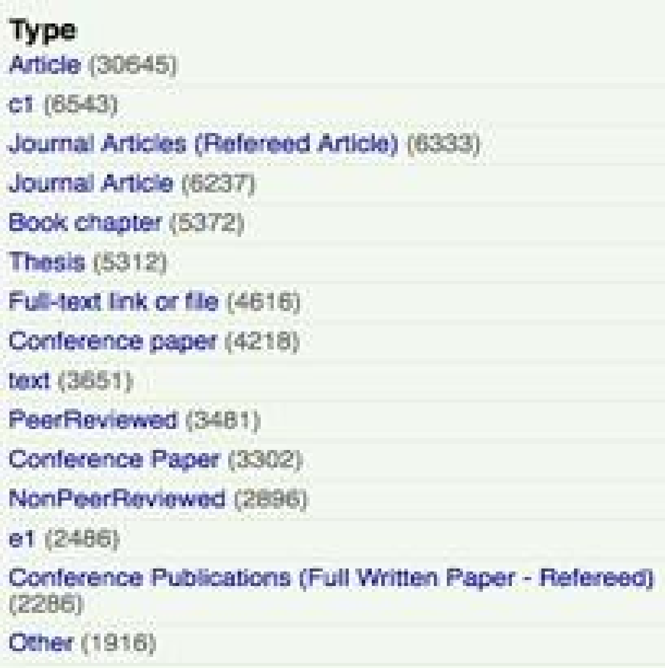
book chapter (13550)

thesis (5864)

report (5521)

more

AUIRC



Type

Article (30645)

c1 (6543)

Journal Articles (Refereed Article) (6333)

Journal Article (6237)

Book chapter (5372)

Thesis (5312)

Full-text link or file (4616)

Conference paper (4218)

text (3651)

PeerReviewed (3481)

Conference Paper (3302)

NonPeerReviewed (2896)

e1 (2486)

Conference Publications (Full Written Paper - Refereed) (2286)

Other (1916)

2.2 XACML

But not all standards have worked out so well. One thing that has definitely been much more painful for the ARROW community than OAI-PMH is XACML, the eXtensible Access Control Markup Language. There has been an expectation that this will be a key standard for repositories but it so far has not turned out that way.

(Actually, it's not even a markup language! Markup is something that you'd put in-line in something else, like a bold tag in an HTML page, or a structural element like chapter in XML.)

Great eXpectations for XACML

XACML is supposed to let you write role-based policies for items in your repository. For example:

> "Only initiated females from the Australian Labor Party and mathematicians are allowed to see this."

But:

- How was your university going to share access policies for mathematicians with my university?

  See the eduPerson spec. Can you figure it out?
- And did we all expect to be interchanging XACML policies. Really?

I was always vaguely worried about how XACML policies were going to work but one day I met Kent Fitch who really nailed it. On the subject of these use cases for XACML where you, an anthropologist want to grant access to a repository to other anthropologists, he asked "What's an anthropologist?[1]"

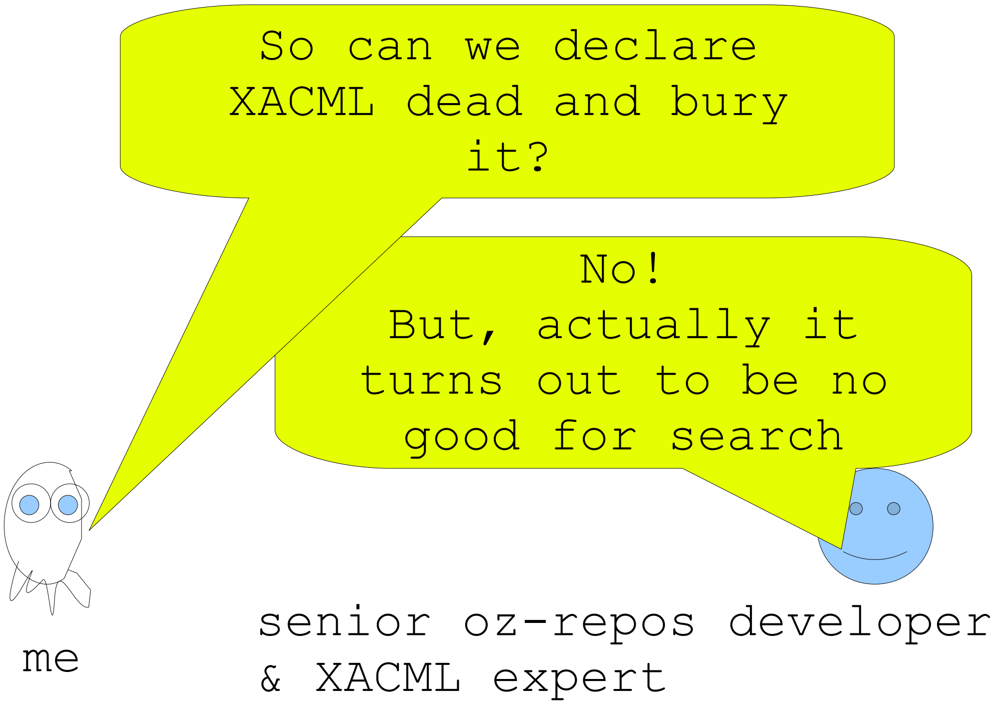
This is a very, very good question. Does an academic working in the education faculty who self-identifies as a visual ehnographer qualify? What if she's got an honours degree in anthropology? In an access federation would the archaeologists who make up our anthropology department count as archaeologists?

Just look at the wide variety of names used to refer to to a thesis in Australian repositories. Remember this is librarians we're talking about here (and maybe the research office), if these information professionals can't agree on what to call something as well defined as a thesis how will they go labeling archaeologists anthropologists, ethnographers, linguists and so on in such a way that I can trust the labels enough to give those people access to research materials in my repository?

But that's not the biggest problem with the unrealistic expectations heaped on XACML.

[1] This was over a beer, and I'm not quoting exactly.

What's wrong with XACML for ARROW*?


So can we declare XACML dead and bury it?
No! But, actually it turns out to be no good for search
me
senior oz-repos developer & XACML expert

Got that? No good for search = no good for > 50% of what you thought you wanted it for.

*I'm not saying there's anything wrong with XACML in other contexts. Might be. Might not be.

The problem is that an XACML policy tells you who has what kind of access to an item, but if the XACML is not able to be integrated with the search index then you can't filter search results per-user without looking at every single item. Very slow, that. But if you don't then you risk letting on that your repository contains something that you should not be disclosing, like the name of a chemical which is the subject of a patent application, or worse the name of a person that you must not disclose.

(The good news is that a project we've been doing at USQ called The Fascinator takes an approach to access control that **is** integrated with search and so far that seems to be working.)

So as I go on and I tell you about how I think ORE is going to be a very useful standard for your repository and the services that go around it, we all need to remember that standards are much less use if there's no possibility of interoperability. For XACML the possibility that you could move an access policy from one repository to another is vanishingly small – remember, we all call a PhD thesis something different. There might be an advantage to learning just one policy expression language but it's certainly not the same advantage as you'd get if you could share policies.

And don't forget that the big problem with XACML is that it turns out to be no good for search. My advice is not to ask for a particular standard because someone else tells you it's important; look for useful software but check that it is going to interoperate with other services and with your own services into the future. Standards do matter.

3 OAI-ORE

Now the the real topic of this presentation.

Officially:

> Open Archives Initiative Object Reuse and Exchange (OAI-ORE) defines standards for the description and exchange of aggregations of Web resources. These aggregations, sometimes called compound digital objects, may combine distributed resources with multiple media types including text, images, data, and video. The goal of these standards is to expose the rich content in these aggregations to applications that support authoring, deposit, exchange, visualization, reuse, and preservation. Although a motivating use case for the work is the changing nature of scholarship and scholarly communication, and the need for cyberinfrastructure to support that scholarship, the intent of the effort is to develop standards that generalize across all web-based information including the increasing popular social networks of “web 2.0”.
>
> http://www.openarchives.org/ore/

What is OAI-ORE?

Open Archives Initiative Object Reuse and Exchange* (OAI-ORE) defines standards for the description and exchange of aggregations of Web resources.

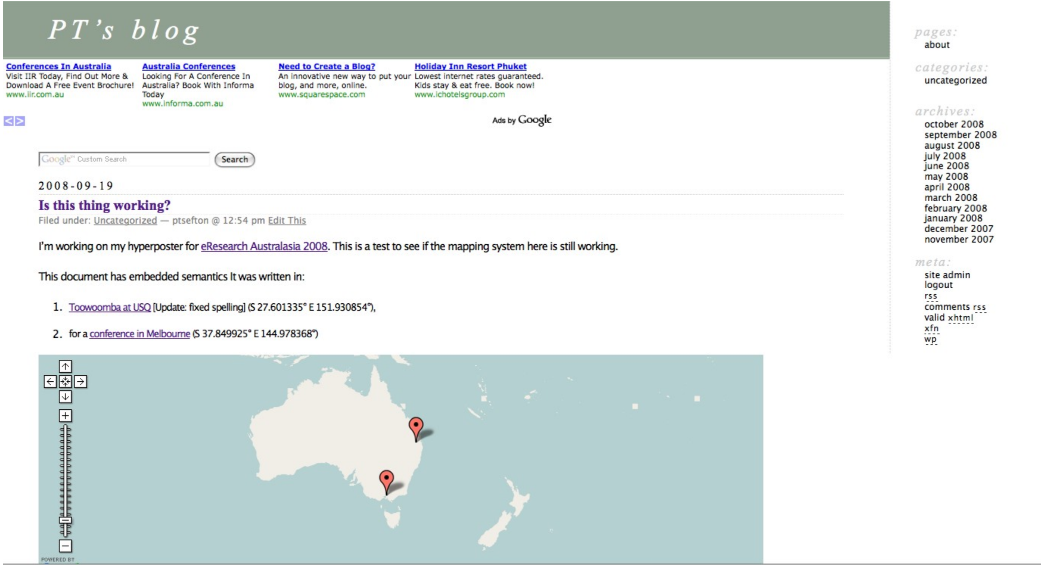
Exchange and subsequent re-use of an object.

Used without permission.

*Shouldn't it be exchange *then* reuse; OER?

Look at my blog. To put content on the blog I use AtomPub – a standard protocol for pushing web content to a server.

Consider a blog post



PT's blog

Conferences In Australia
Visit IIR Today, Find Out More & Download A Free Event Brochure!
www.iir.com.au

Australia Conferences
Looking For A Conference In Australia? Book With Informa Today
www.informa.com.au

Need to Create a Blog?
An innovative new way to put your blog, and more, online.
www.squarespace.com

Holiday Inn Resort Phuket
Lowest internet rates guaranteed. Kids stay & eat free. Book now!
www.ichotelsgroup.com

Ads by Google

Google™ Custom Search | Search

2008-09-19

Is this thing working?

Filed under: Uncategorized — ptsefton @ 12:54 pm Edit This

I'm working on my hyperposter for eResearch Australasia 2008. This is a test to see if the mapping system here is still working.

This document has embedded semantics It was written in:

1. Toowoomba at USQ [Update: fixed spelling] (S 27.601335° E 151.930854°),
2. for a conference in Melbourne (S 37.849925° E 144.978368°)

pages:
about

categories:
uncategorized

archives:
october 2008
september 2008
august 2008
july 2008
june 2008
may 2008
april 2008
march 2008
february 2008
january 2008
december 2007
november 2007

meta:
site admin
logout
rss
comments rss
valid xhtml
xfn
wp

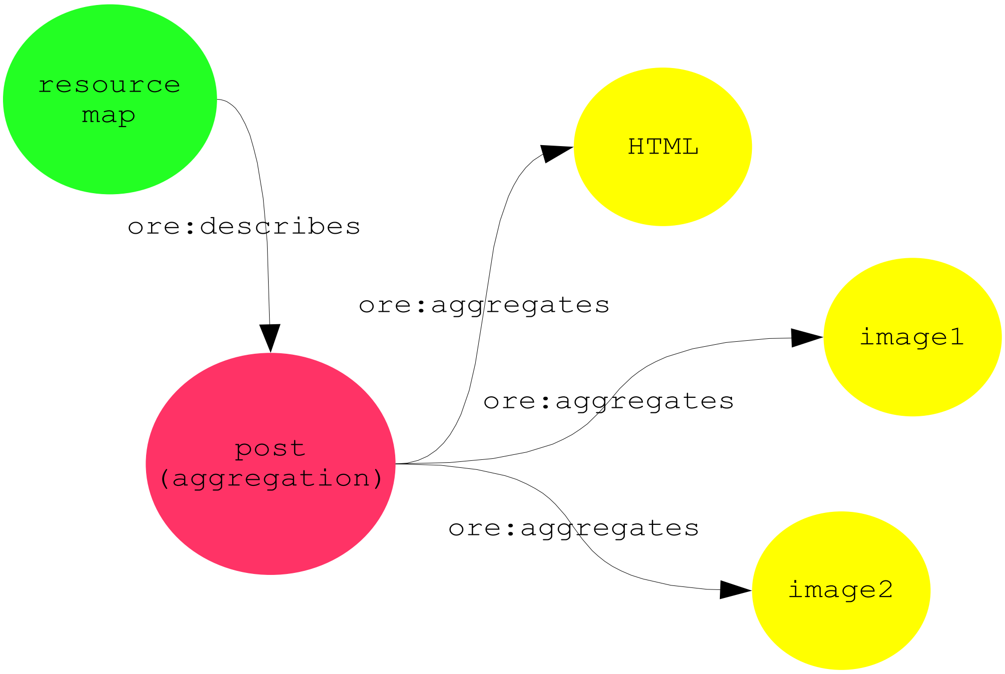
Which bits of this page are part of the 'post'?

One of the really dumb things about blogs in general is that to the blog *application* an article usually consists of just the HTML part. In WordPress, for example a post is a bit of HTML and some metadata. If I want to include a picture it has to go into an uploads directory. To the reader, it is perfectly reasonable to think of a blog post with a couple of images as a single document but most blogging software and the ATOM protocol don't really support this. To further complicate matters a blog could be laced with advertisements and adorned with web 2.0 widgets that most of us would not consider part of the content. In the example above there's an automatically generated map which is actually **not** part of the content I typed in but which is generated from it.

ORE offers a web-based way to describe an aggregate resource, what I think of as a post, made up of the HTML part and the images that the software likes to think of as uploads but which I think of as part of my document.

Only it doesn't work like that. Not for WordPress. The plugin I installed at my blog doesn't treat stuff from the uploads directory as being part of the post. (But it will, it seems).

The ORE view of a blog post



The ORE tutorial starts off with simple stuff like this, but it quickly gets fiendishly complicated, and you have to deal with prose which is obviously very carefully worded but can be quite impenetrable.

More about OAI-ORE

- You can decorate the simple ORE graph above with lots and lots of extra metadata and rich semantics.
- The aggregation is *imaginary*. It is not the same as the resource map. It will have a URI something like:

  http://ptsefton.com/2008/09/19/is-this-thing-working.htm#aggregation

  This imaginary aggregation can spark long discussions.

  I can't explain it to you.

  See the list.

I'm looking forward to a time when there is some consensus on a 'blog post' type I can attach to the aggregation (failing that we can use NLA-style normalization a la ARROW discovery).

What are other people doing with ORE?

- At the Repository Challenge at OpenRepositories 2008 some UK developers used it to move repository content between ePrints and Fedora.
- There was a recent competition, the ORE Challenge at RepoCamp 2008 in which people presented various ORE-based ideas.
- People are writing software libraries to make it easy to use.

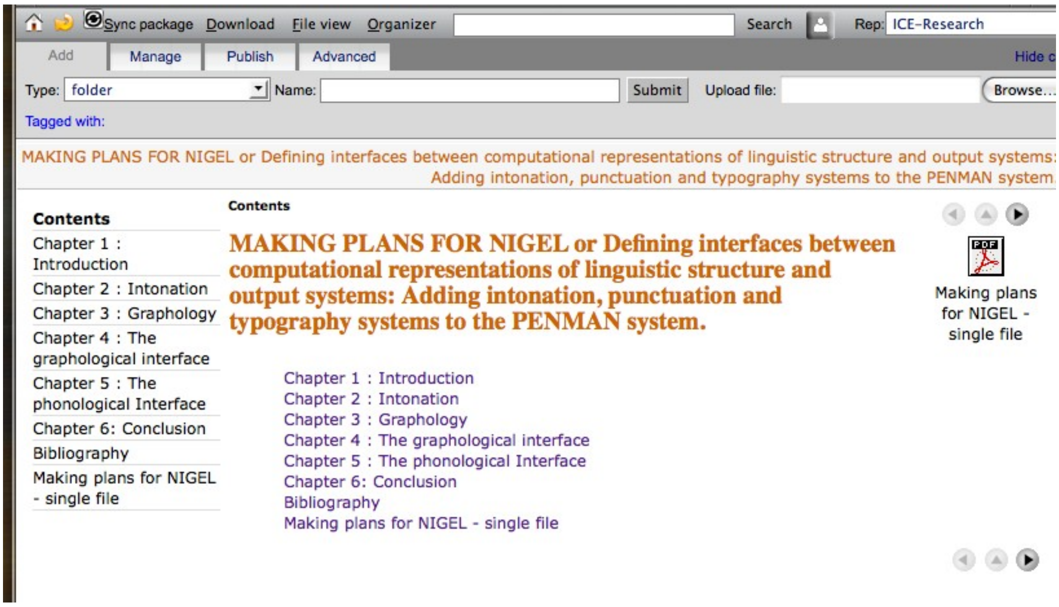
(I don't much like the entry which won the ORE challenge, which was just a worse way to visualize relationships than you could get using text).

What are we (USQ and friends) doing with it?

1. Pushing theses and other research content around, as part of the JISC-funded ICE-TheOREM project.
2. Harvesting repositories into The Fascinator.
3. Ingesting image collections into The Fascinator, exploring possibilities for working repositories for eResearch.

Here's a thesis in the ICE content management system.

A BA (Hons) thesis in ICE



Rep: ICE-Research

Type: folder

MAKING PLANS FOR NIGEL or Defining interfaces between computational representations of linguistic structure and output systems: Adding intonation, punctuation and typography systems to the PENMAN system

Contents

Contents

MAKING PLANS FOR NIGEL or Defining interfaces between computational representations of linguistic structure and output systems: Adding intonation, punctuation and typography systems to the PENMAN system.

Making plans for NIGEL - single file

This is a complex object with hundreds of HTML and PDF and image files.

Using ICE we can push that thesis over SWORD using ORE to map it, to ePrints, or Fedora, or (soon) into the VALET system.

4 Conclusion

But what will ORE do for me?

IMHO here are a few examples:

1. Replace or supplement OAI-PMH for **moving content** between repositories – not just harvesting metadata but even upgrading to new software.
2. Improve research **tools like Zotero** by making it easier to tell the tool what to download when saving a local copy of a paper.
3. **Replace the use of METS packages** in work like APSR's OJS to DSpace demo. [I may be on my own here]
4. Allow for **thesis by publication** in a very elegant way.
5. Pave the way for a new repository architecture which **understands content models**. (No more discussion of *atomistic* versus *compound* objects).

Finally

To help ORE it along I suggest a minor rebranding. From now on, call it...

ÖЯ3

http://en.wikipedia.org/wiki/Heavy_metal_umlaut

http://en.wikipedia.org/wiki/Faux_Cyrillic

http://en.wikipedia.org/wiki/Leet