Towards (Australian) repository interoperability using OIA-PMH

Jim Downing tagged the presentation I posted on Tuesday What the OAI-ORE protocol can do for you as "Apart from the ORE parts, this contains a nice exposition of the difference between standards and interoperability."

That tag nails a lot of what the talk was about. ICT Standards are nice, but they don't always guarantee interop. The main example I gave of this was the National Library of Australia's ARROW Discovery Service. It creates a normalized view of what's in Australia's institutional repositories. But the half-finished harvest we're doing with USQ's Australian University Institutional Repository Census (AURIC) shows the underlying chaos, more politely described as "the diverse range of ways repositories describe their content".

The reasonably rosy picture you see at the ARROW Discovery Service is is not a result of true interop and I fear that it is not scalable. Instead it depends on a great deal of ongoing work by the maintainers to keep all the normalizing rules up to date.

There are other situations where data need to be moved around where the recipient of the data is not going to patiently and laboriously normalize the your ad-hoc content. Lets digress to look at one: the forthcoming ERA for Australia, which replaces the never-ran RQF.

I think I'm right in saying that the Australian Government is not going to be particularly understanding if your institution submits its data using a local 'standard'. If they think that a particular kind of research output is of type journalArticle then you won't be able to submit something called "Article, Journal".

If you want to use the repository of research outputs to feed a report to the government system then you will have to source or create some kind of report, or adaptor or something. Contrary to the fears of at least one repository manager I have spoken to recently this doesn't necessarily mean that you have to change what's in your repository (unless the metadata you've been collecting is inadequate to make the required distinctions). But it will mean that some bit of software needs to be created to provide the reports that you need.

I think what's causing some stress is that our repositories don't really have adaptors in all the places they need them yet. We don't have anything like those little power adaptors that let you plug European appliances into Aussie wall-outlets. (By the way, Rick Jelliffe has a great post on interop relating power points to office document standards.).

Returning to the ARROW discovery service, the available-on-request guidelines say (very reasonably):

> Populate as many fields as possible. It is a good idea to populate the Type element at all times, and to use the MACAR list to do so.

But there is no straightforward way using standard IR software to both use a MACAR type for the external view of a repository, vs the internal view. At USQ, for example the official university nomenclature for theses and dissertations is at odds with the MACAR types.

So, here's a three step process that I think could move us towards better interop. that **could** be made more sustainable and less of a drain on the NLA:

1. The NLA **publish all the rules** they use to normalize repository content in a public place. This might not be in the form of a standard, but it can be made human readable.

   At USQ we've been adapting NLA rules in our work on The Fascinator and with the AUIRC, and they're going to be in The Fascinator's distribution as an example.

2. The repository community **works to create adaptors** for the various repositories so that they can move the normalization closer to home. Instead of feeding whatever you happen to have to the long-suffering ARROW discovery service you take responsibility for mapping your local view of your data to the shared standardized view.
3. **After a testing period** the NLA switch over to the new system and start rejecting out-of-band input.

This is of course, just an example for discussion, to illuminate the general issue of interoperability. I don't set NLA policy. It's not my service.

There are lots of ways this normalization could be done, but how about using an OAI-PMH **adaptor**. Hook one side of it to your repository and expose the other side to the harvester. It would sit there quietly normalizing the content.

It doesn't matter where the adaptor runs, but it **does matter who looks after the normalization rules**. It's not sustainable to expect staff at the consumer end to keep adapting to non-standard inputs unless there is a very clear case for ongoing funding for such activities.

Being a techo I think of this normalizing OAI-PMH adaptor as a proxy and it turns out that a proxy is actually one of the recommendations from:

> a small pilot project to set up a European doctoral e-theses Demonstrator. The work was funded by the Joint Information Systems Committee (JISC) in the UK, the National Library of Sweden and SURFfoundation in the Netherlands. The project has been performed by SURFfoundation.
>
> http://www.surffoundation.nl/download/ETD_LessonsLearned_Full-Report+Annex.pdf

It mentions proxies:

> The best solution is to create national proxy services. These proxies are gateways that on one side harvest the metadata from local Institutional Repositories, and on the other side have an OAI-PMH gate that can be used for service providers to harvest the national proxy. The advantage of a national proxy is that it can centrally normalise the metadata, and deliver unambiguous metadata to service providers.

The ARROW discovery service is already a national proxy – it normalizes and acts as a gateway - but I am concerned about the centralized maintenance of the normalizing rules – there's not much of a disincentive to changing stuff around if someone is going to clean up after you, is there?

From the same report, the tedium of normalizing:

> This has to be done for every type of field of every repository, which is a labour intensive enterprise when offering a high quality service. This is a short term solution, the long term solution will be that repositories offer standardised content that can be easily used for interoperable services.

Right – what I'm suggesting is finding a way to give the NLA proxy/normalizer back to the providers. One way would be to make software available that lets people adapt their OAI-PMH feed without having to change their repository. Or the NLA could make the rule-sets editable by staff from the sites they harvest.

Given the current state of the repository art I would not propose that a harvester service suddenly start enforcing standards (ie rejecting non standard input). But, if we could come up with some freely available adaptor infrastructure then would it be reasonable for the NLA to, after a giving people time to adapt, stop accepting ad-hoc data and stop being responsible for making the service consistent? (That's a genuine question. Would it? Use the comments below if you have an opinion).

There are other models.

The Zotero research tool project, for example relies on volunteers writing vast numbers of adaptors to let Zotero harvest metadata from various kinds of web sites. I don't see anybody seriously suggesting that all this work is pushed back to the sites. I wonder about opening adaptor rules up to everyone wiki style. What would happen?

If it turns out that the ARROW discovery service is not all that important to the owners of the repositories it harvests then there's no use expecting them to take over responsibility for normalizing their outputs. In that case it's either it's worth supporting at a national level, via a community or it's not worth having. Repository owners – do you value the service?

Finally, a reflection on the process I went through in writing this post. What I thought I was sitting down to write was a survey of potential ways to share and reuse the normalizing cross-walks that have been developed at the NLA. But in the process of thinking it through I came to the conclusion that a normalizing OAI-PMH adaptor is a pretty good idea. Then I realized that the Arrow Discovery Service is already doing that job. I was thinking about all sorts of complicated machinery but now I think that we have nearly all the technology we really need, it's just a matter of distributing it the right way.

Returning to Jim's perceptive comment re standards versus interop; the OAI-PMH Standard is more or less working for IR's in Australia but the interoperability is limping along. I think we now have a great opportunity to think about what interoperability is worth to us as a community and put the power into the right hands to make it happen. Should it be central, at each repository site, or crowd-sourced? Choose one or more, or give up.

I'm very interested in this as a test case for how we handle content models in OAI-ORE, if we want to start swapping theses and journal issues and the like are we going to be able to get working interop from the start or will we need to go through the same kind of process as we're going through with OAI-PMH?