
Trove is an aggregator. Every Australian University repository contributes directly to Trove. Some other research organisations and tertiary institutes do as well. The actual collections vary – most universities add theses and publication outputs from their academic staff. Some add data sets and, others the open journal systems they host. Course work collections, digitised collections, there’s significant variation in the kinds of collections that universities put into Trove.

As you probably know Trove had to put our aggregation activities on pause mid-last year. Hopefully you’ve also heard that, thanks to a very passionate community of users and supporters many of which are in this room, the Commonwealth Government provided Trove with very welcome 4 year modernisation funding before Christmas. This means we started harvesting again this March. Over the 9 months we weren’t harvesting we acquired a long list – a backlog of maintenance work on repositories like yours where people had naturally come time for upgrade or moved to a new system.

When we started harvesting again we’d had 9 months off. We were coming back with fresh eyes. Which was a good thing, because it’s a good time to re-evaluate what’s going on with data standards.
This afternoon I’m going to talk about a few of the standards and identifiers that are out there, discuss what Trove is seeing from Australian Institutional Repositories as a whole and briefly touch on what we should be doing going forward.
There’s been a proliferation of standards over the last few years to help track researchers, funding grants, publication outputs, licensing and open access compliance.

We know that these standards are being implemented in Australian repositories.

At the Trove end, we see incredible variation from one repository to another in what data they’re able to share.

This morning’s discussion on ORCID alone shows how ubiquitous that one service has become.

And yet, looking at the data we see coming into Trove, much of it doesn’t contain ORCIDs.

It didn’t seem to marry up to what we know is happening in repositories.

So what’s going on?
Here are the shared identifiers that have come to be pretty much agreed on in the five years I’ve been working with Australian repository data:

**ORCID**

Why? We’re all agreed that implementing ORCID brings value to the repository that far exceeds the cost of implementing. They help us track research outputs which can feed into reporting activities, which can influence future funding and so on. They can help us track students as they become early career researchers and over the course of their career. We can link funding and research effort to actual output for our institutions. Grant identifiers, particularly for the ARC and NHMRC. Again, reporting, very useful. They help quantify the link between funding and effort to published output. They can help track compliance with the ARC & NHMRC’s open access mandates. They help the Australian public pinpoint the research their tax dollars have funded.

**DOIs**

These are an oldie but still a goodie. They’re being used to track a lot of linkages. They can track use and re-use of a publication, as well as its impact over time through reverse engineering of citations. All these identifiers give us a better view of the research process and how people, funding and output are linked. As a side benefit, they give us the linked data to provide the promise of fantastic discovery to end users.
That’s the common set of identifiers, what about our common data standards?
Everyone gives Trove simple Dublin Core, it’s a basic, baseline metadata record sharing standard.

A few additional standards are relevant in this context:

The NISO access and licence indicators recommended practice is probably the most applicable.

It came out in 2015. They exist to help us track, find and report on what material is open access. It makes it possible to interrogate whether open access mandates are being met. Also help maximise the re-use of existing material by making it easy to search for and find open access material, as well as material that meets certain licence conditions.

The other shared standard seems to be the support for use of Creative Commons licences. Creative Commons and Rights Statements.org are two licencing standards that make it easier to find and re-use open material.

So we have great universal identifiers, and we have great standards to put with them. They’re making life much easier within individual repositories. Are they making a difference at the national level?
ORCID uptake is really clear in Australia.
In March 2015 when I first went on the hunt for Australians in ORCID the process was able to identify 2,353 ORCID profiles and I thought that was fantastic.

Today there are 17,603.

Those are researcher profiles in our people zone.

That's fantastic.

People are signing up for ORCID, they’re putting enough information into their profiles to identify them as Australian and Trove is able to programatically locate them.
So when you’re searching Trove’s People zone you’ll have these Australian ORCID profiles mixed in to find. And as a sector we can also get some insight on what is happening with Australians and ORCID. See on the left hand side under that ‘Occupation’ facet that Epidemiology is the #1 field of Australians in ORCID.

And that health, science and education are the fields where people are signing up to ORCID.
There's also this nice ‘Related to’ facet on the left. It lists every Australian university, and how many researchers Trove has been able to relate to them.

If I click University of Melbourne up the top there I’ll see all people who studied or worked at the University of Melbourne, signed up for an ORCID, that we’ve been able to identify. Every Australian University, plus the ARC and NHMRC are present in that facet.

So on the “getting people to use ORCID” front we’re doing pretty great.

They’re putting their organisation information in too, which is going to help us track research outputs and students over their careers.

What about doing it the other way?

What about linking publications to authors? How’s that going?
Not so great.

Australian University repositories add 1.9 million records to Trove. Just under 13,000 contain an ORCID, less than 1%. There are less publications than there are people with ORCIDs.

While people are signing up for ORCIDs, making profiles for themselves, hopefully adding their publication data to those ORCID profiles - they’re not adding their ORCID to the repository metadata record for their publication. Or if they are, that data is not being output from the repository to Trove.

We’ve got a long way to go in terms of publications linking back to the associated people.

From what I can see in the data that is coming in, two universities are doing this really well.
University of New England and University of Melbourne.

The database vendor GALE also has some ORCIDs coming through.

That might signal the start of the automated metadata dream, where authors include their ORCID at submission time and it tags along with the metadata all the way through the metadata record’s life.

This graph could be a bit disheartening, but I think it signals we’re still at the start.

ORCIDs are being implemented in local systems but not yet being output with other metadata.

And here’s why I think that...
ARC and NHMRC grant identifiers.

In March 2014, Paula Callan (QUT), Mark Gregson (QUT), Kerrie Burn (ACU), Tony McCall (ACU) put out the guide to tagging institutional repository records related to ARC/NHMRC grants.

This graph shows the number of publications that have come into Trove with those grant tags in the intervening three years, which is fantastic. Started off a bit bumpy when a couple of repositories led the way. And you can see there are fits and starts as another repositories start adding those tags.

Now we’re smoothing out, growing at a steady rate and about to hit the 20,000 mark for ARC-tagged publications. Which says to me that grant tagging is becoming common practice and this standard is starting to be widely implemented across the sector.
Compare that to this graph, which is researchers we’ve been able to find in ORCID, who nominated they had received an NHMRC or ARC grant. It’s a bit flat and bumpy still which confirms for me it’s still early days in sharing ORCID data beyond the actual people records.

DOIs are the third identifier I mentioned. They’re old school and pretty ubiquitous at this point. More than 750,000 of them in Trove. The problem with DOIs is their form.
Recommended form has changed a bit over time. Here are just some of the examples of the form they take in Trove:

- Just the number
- DOI in capitals, colon, space, then the number
- DOI in capitals, colon, no space, then the number
- Doi in lower case, colon, then the number
- As a URL with http and the dx prefix
- As a URL with http and no dx prefix
- https with the dx prefix
- https with no dx prefix.

At this point in time CrossRef recommends using the final one, and Trove definitely prefers a URL. This allows us to do linking and data enhancement.
Where to from here?

Australia is lucky in that we have a history of collaboration going back decades in the Library sector and more than fifteen years as IRs.

We already have this infrastructure and practices established that allow us to do data sharing and analysis and to build collaborative tools because we already aggregate into one big data set. To realise that potential, we need to keep working towards consistent data, especially as some of our systems are quite mature.

So it could be time for a repository health check, particularly on the metadata you share. I suggest checking the following identifiers:
Make sure you’re putting out your repository URL, and that’s in the dc:identifier field. That’s a basic one but we do see quite a bit of variation.

If you have ORCIDs, put the full URL form and put it in a dc:relation.

If you have ARC or NHMRC grant identifiers, put them in a dc:relation in the full URL form as well. And the same for DOIs.
It would be worth doing a checkup of some of the other standard practices as well.
How are you including open access indicators? Do you use the 2015 recommended practice, or something you implemented locally before this recommended practice existed?
Are you actually exposing which items are open access and which aren’t?
If you can determine what is and isn’t open access, you should expose that. In Trove we use it to power that green ‘View Online’ tick.
If you have Creative commons licences, do you put textual statements? Is it possible to put machine readable ones as well, and by that I mean URLs.
Would RightsStatements be useful for your repository and if so how would you store and share that data?
The best time to do this kind of check is when you have an internal change, upgrade, or migrate to a new system.
Check your OAI feed to make sure that any new data is also being exported as well.
I’ll leave you with the message that if it’s been awhile since you last updated your OAI output, it might be time for a health checkup.