Today I’m going to speak about Trove as a major consumer of Australia’s machine-friendly repositories. And the benefits of being a machine-friendly repository in turn.

Here’s how my half an hour will go this morning:
I’m hoping most of you have heard of Trove, in case you haven’t I’ll start us off by racing through a quick overview of what Trove is.

I’ll talk about the idealistic dream around aggregation, and lift the lid on the messy reality.

The benefits of aggregation to users, to repository managers and to Australian society.

The benefits we’ve experienced of being really open and machine readable. There are benefits to ourselves and to our users.
What is Trove?
It’s both an aggregator of metadata from Australian repositories, and its own repository of digital content.
Trove has four basic functions:

It’s an aggregator of collections from research, cultural heritage and community organisations, as well as Library vendors. Range from big universities and national institutions down to small cricket clubs and volunteer run museums.

It’s a digital collections host – most known for digitised newspapers but over the last three years has expanded into hosting other types of digitised and born digital content.

Platform to research and build – some basic tools in the interface and a number of APIs to make the data freely available.

Community end users. They’re the ones who do the research, make connections, correct and enrich the data for others.

Those are the four basic functions of Trove.

I’ll start off by talking about that role as an Australian aggregator of metadata.

There are different ways to get metadata into Trove:
Some are bulk transfers like the tram, and some are steadier streams like the pedestrians:

- MARC records via Libraries Australia.
- Select databases hosted by vendors GALE and Informit.
- A few foreign fulltext repositories including Hathi Trust and Open Library.
- Direct-to-Trove – this is harvesting from Australian institutional repositories and data stores.

A decade ago, a bunch of format and subject based aggregator services had sprung up in Australia, seven or eight of them hosted by the National Library alone.

These brought together resources from open repositories into a single place users could search. Australian Research Online aggregated theses and academic outputs from institutional repositories at our universities.

Picture Australia aggregated images of photographs and art works from libraries and galleries.

Music Australia brought together sheet music, sound and biographies from music libraries and national institutions.

Australian Newspapers was a service that hosted collaboratively digitised Australian newspapers.

Fast forward to 2017, and those separate services are gone - aggregation looks completely different and there’s a much wider group of organisations we work with and a wider spectrum of collections.

Very different back room processes.

There’s often a tension about centralised vs distributed models. Do you host everything in one place, which offers a convenient single place for users to search? Or do you host everything in its own repository, which means each collection can be shown in a system built to best highlight all its unique attributes?

Trove is a mixture of both. It’s a repository where collaboratively digitised and born digital material lives.
Items like this digitised newspaper – you see it within Trove.

And also an aggregator of content from other Australian collections.
Which means we allow users to find items like this photograph in Trove. We show a description of the item, plus a thumbnail if the user is lucky. But send them off to another organisation’s website to view the full item. Trove has collections from about 1,000 Australian organisations. We harvest 400 collections from Australian repositories, representing about 250 organisations, via the Direct-to-Trove pathway.
In a dream world every repository would be “machine friendly”.

Which would make aggregation all heavenly butterflies and harps.
Every repository would use standards and stick to them.
They’d implement controlled vocabularies.
And global identifiers.
Metadata that arrived in different formats like Dublin Core, and MODS, and MARC, and RIF-CS would be so easy to translate between, using freely available online crosswalks.
With standards, an aggregator like Trove would be able to automatically link data across contributing repositories and provide a fantastic inked data user experience based on really rich data.
Every resource would be online.
Freely available to the whole world.
With clear re-use conditions.
The more collections you can add, the better.
That is the dream of open data and open repositories from an aggregator’s point of view.
All we would have to do is point Trove at a new collection, and as if fluttered in on Angel’s wings, that new collection would appear in Trove.
The reality is nothing like that.
There seem to be more vocabularies than terms, and some organisations describe using free-text. There are not one or two but lots of different standards that are used and even when the same standard is used, it’s implemented differently at almost every location.

Every content partner we work with needs some normalisation of their data before it comes into Trove. And if we have to crosswalk from another format, it has to be tailored to the single organisation’s implementation.

Different standards don’t always come together nicely. For example, our Libraries Australia MARC records don’t always marry to records from the vendors who contribute their journal article databases. There are difficulties linking aggregated content with the digital content that lives in Trove, even when they represent the same physical object or series.

The Bush Fire bulletin is a historic journal that’s been digitised into Trove, in partnership with the NSW Rural Fire Service. They now contribute their more recent, born digital issues as well. And we’re not yet able to provide one continuous service, where someone can browse through the digitised versions and then on into the born digital ones.

Servicing a wide range of organisations means a wide spectrum of capabilities at the content partner’s end. A university might have a whole team of information professionals to manage an institutional repository. A small historical society can have a single volunteer a few hours a month. For us, that means we need to be capable of providing really different kinds of support to suit the organisation.

We aggregate across cultural, research and community sectors – libraries, galleries, archives, museums, universities and research repositories, government agencies and departments, cricket clubs and RSLs and local historical societies. Part of it is that the metadata we’re aggregating describes such different resources.
Just within the university sector we’re likely to get:

- theses
- datasets
- published academic output including research papers, journal articles, book chapters and pre-prints
- coursework
- open access books the university publishes
- open journals they’re hosting.

Trove might get other bits of data not directly from the IR, but managed by the same team in some way.

One of the advantages of open data sets and open repositories, is that a public-facing aggregation service like Trove can bring those data sets together, even when they’re hosted by separate organisations, to provide a more connected user experience.
Examples of that include:
Data set descriptions that come via ANDS RDA.
Biographical records. In the uni sector we try to source biographies of Australian researchers from ORCID.
Also get a smattering of small RedBox MINT collections of identities or names.
Then there are other parts of universities that contribute to Trove in their own right.
Examples include:
- museum archaeology and specimen type teaching collections
- subject specific repositories with working papers
- university archives
- heritage collections managed by the Library or someone else
- outputs from research projects like AusStage.

Now all of those are just within the university sector. Broaden out a bit and two thirds of direct-to-Trove content comes from other repositories with items like:
Local history photographs
Objects from museum collections
Archival material like letters, manuscript collections
Maps
Biographies
Political press releases
Radio broadcasts – both the written transcript to search across and the podcast to listen to
Art works
And many more kinds of resources I can’t fit on this slide.

The point is, we deal with lots of different data types, lots of different formats, and lots of different systems.
We do a lot of data normalisation.
This is an example of some of the basic steps the Trove team goes through to add a new collection to Trove.

Some of the systems we aggregate from are built to share, most are built to search and display online within their own system – quite beautifully I might add.

But data sharing is often an after-thought and there may not be capability within the organisation to facilitate it.

I'll be the first to put my hand up and admit data-sharing was originally an after-thought for Trove as well.

One other reality we’ve had to face, is that just because data is open, and resources are relatively easy for us to interoperate with, doesn’t mean they’re right for Trove.

Not a problem I ever thought we’d have.

But there are collections that users actually don’t want in Trove.

Just to re-emphasise, Trove is a place that people come to find Australian resources. In the early days of Trove we hadn’t quite worked that out, and included some foreign open repositories of data.
Including OAIster, does anyone remember OAIster?

It was/is content aggregated from world-wide OAI-accessible repositories.

We added OAIster data to Trove back in 2009 when we thought that the more online freely available resources we could help people find, the better.

By including aggregated metadata from all over the world, we were helping point people towards resources they wouldn’t otherwise find.

But more and more we were getting complaints from users about foreign content and how it created noise and stopped them finding what they wanted.

They come to Trove to look for Australian content, and then have to wade through pictures of buffalo roaming mid-western prairies and couldn’t figure out why. Isn’t Trove Australian?

Users didn’t say “OAIster”, because they didn’t realise what they were looking at.

But we would get enquiries saying they couldn’t find Australian images.

Or that an item had a dead link and why couldn’t we tell them where it actually was?

We didn’t have a pathway to regularly update the OAIster data, so not only was it foreign content, it also became stale.

That’s really bad for a platform where one of the main attractions is freely available, accessible, online materials.

This data even contained secondary copies of records we were also getting directly from Australian repositories, except this secondary OAIster copy of the record had an old dead link and the repository manager couldn’t figure out why.

So in the last few weeks we’ve removed that OAIster content from Trove altogether.

And are already getting happier users.

Which I guess is a good lesson – more is not always better, sometimes it actually gets in the way.

Especially if the service you’re running has a particular speciality or focus – in our case Australian – and you’ve got content that doesn’t fit with that.
The big benefit of aggregation is:

Homogenising all these different types of content from lots of different Australian Institutional repositories alongside the digital content that lives within Trove. The point is to help people make connections between all these items and tell stories.
This is one of them.

The story of Jessie Webb was put together by members of the Trove team, who were on the hunt for a good Australian story. They started off looking at the biographies in Trove, browsing the list of occupations to select someone who would have an interesting, visual and adventurous story, that would also be covered (out of copyright!) by collections already in Trove.

Jessie was born in 1880 in New South Wales. They found significant dates (birth/death, city visits, travel, awards) from digitized newspapers across the country. Used extra research from books, diaries and photos to shape an understanding of Jessie’s life. Including a published book on Jessie’s Capetown to Cairo trip, information from her diary of her travels, and photos of each location such as sketches of Ellerslie Station.

Filled in extra surrounding facts from a state library, a university archive, and a museum collection from different parts of the country. Having a single place to search in Trove makes life convenient. It’s the ability to make connections and tell stories like Jessie Webb’s by bringing together the resources from lots of different organisations that is the real power in Trove. And that’s made possible because of open, machine friendly data from hundreds of Australian institutions. Research rarely relies on one resource.
There are lots of benefits of having openly accessible data, of being machine-readable and human accessible.
Trove has a number of APIs, plus tools in the user interface, to help maximise the value of the data in our repository.

Probably the most significant internally is that we are massive consumers of our own data.

Earlier this year we had a new KPI introduced – regularly report on how many works we get direct-to-trove, by sector, and what % are online, and how many organisations do they come from, and how many are libraries.

The edict was that we need to produce this every month.

To produce it requires laboriously going through Trove to manually run hundreds of searches. Or does it? Well, no...
We were able to knock out this, granted ugly but very effective, PHP script. It sits completely separate to Trove, and accesses the Trove API just like any member of the public does.

It breaks down #s on how many works we’re getting by sector.

And then further breaks down the kinds of resources we’re getting by item format.
Which lets us get insight into our own data, without requiring any IT resources or changes to our own service.

Sure it’s not all that to look at, but for a few hours work we’ve now got reliable regular data and we can use other tools to produce something a bit prettier on an irregular basis where we need more than numbers.
Like this, which took about five minutes to produce in Excel from our regular script. We also use our own API to do data quality control.

Trove is quite a complex myriad of systems underneath the hood but the API presents them coherently.

When we want to do something like update ORCID biographies...
We get the records out of our own API, enrich it with data from the ORCID API, then feed the updated biographies back into Trove via our Harvester.

When we want to put together a list of the digitised fulltext journals Trove is now hosting, we harvest that data set out of our own API and use Microsoft Excel to give us a unique title list. Having our own open data API has also been a great way to upskill staff. They have to support public users of the API, they have to be able to do this kind of data analysis, so they’re encouraged to access the API as any other member of the public would, and experiment.

Those are our internal benefits...
We know that exposing our open data also benefits certain groups of users. One of those is academic researchers.
This is a tool called QueryPic. It was built by Tim Sherratt, on the back of the Trove API some years ago and is freely available. It allows anyone to type in terms and it will produce a graph, plotting how often those terms appear in the Australian newspaper corpus over time.

In this example I’ve plotted three terms, “digger”, “Anzac” and “Aussie”. Unsurprisingly, the term “Anzac” comes into being during WW1. “Aussie” is a term that doesn’t come into use until the end of WW1. And “digger” – a term often used to refer to Australian soldiers – actually comes into use in 1851. Which this graph really clearly shows. And would cause a researcher to probe deeper, why was that term in use before WW1? They’d soon discover the term was used to refer to people who were digging for gold in the gold rushes.

Tools like this can expose events and changes in society really quickly, in a way that browsing through a list of search results in a user interface can’t. This is a tool we never would have thought to build. By keeping our data open, freely available to everyone, we’ve enabled researchers to build the tools that suit them. And we know that other researchers are using this tool to interrogate their own areas of interest.

I read an interesting paper coming out of the University of Queensland two years ago about tracking the emergence of surfing in Australia, and challenging pre-conceived notions of women’s uptake of the sport. It was based on this QueryPic tool. I also know of a current PhD student at the ANU who is researching the history of the musical life of early twentieth century Australia, with a particular focus on the Sydney Symphony Orchestra as a modern cultural institution. She is using QueryPic to look at the prevalence of the terms ‘recital’ and ‘concert’ in Sydney during the first half of the twentieth century.

This tool isn’t part of Trove. It’s because Trove is machine friendly and exposes this data for other people to re-use that this new kind of great research is being done, and challenging pre-conceived notions of our own history. Trove’s open data is good for the Trove team, it’s good for academic researchers, it’s also good for creative people.
This is a tool called Culture Collage built by Zenlan Ltd. It takes images from Trove and displays them in a different interface. In this particular example, it’s a Trove search for fireworks.

Culture Collage is freely available online in its own right. It offers different ways for people to engage with Trove data outside our own interface.

Within our own sector, Library suppliers are using Trove to enrich their own offerings. I know of at least two companies who access the aggregated Trove theses articles and include them in the discovery layer products they offer to Libraries. Enriching the vendor’s product, and making the theses even more widely discoverable.

There’s also untapped potential to build analysis tools that do comparison and statistics across sectoral data.
For example what kinds of works are universities giving to Trove?
Which state do works come from?
How many Australian Theses are digitised and online so far?

There's potential to evaluate initiatives and the power is open to anyone, not just the staff at the National Library.
Trove’s experience is that being open returns a benefit.

If you want to quantify the benefits here’s a monetary one:
We estimate that Volunteers have spent 597 years and an equivalent of almost $40 million worth of volunteer time to correct newspaper text.

An Australian woman currently has a life expectancy of 84.5 years (http://www.aihw.gov.au/deaths/life-expectancy/). Which means that 7 Australian women, working from the day they’re born until the day they die, correcting text, has been voluntarily given to Trove.

Work we could never resource.

I’ll leave you with that thought.
Program for event at:
https://www.conftool.net/or2017/index.php?page=browseSessions&form_session=257