Easy as ABC – A triumph of re-usable metadata

By Julia Hickie and Mark Raadgever, Trove Support Team, National Library of Australia

Introduction

At the end of 2013 the Trove team at the National Library of Australia embarked on an exciting project to bring Trove's current affairs coverage into the twenty-first century. The Australian Broadcasting Corporation (ABC) Radio National (RN) website exposes a wealth of contemporary content on cultural and political life in Australia. We knew that if we included these resources in Trove we could give users a current affairs discovery experience starting with the first Australian newspaper printed in 1803 and continuing all the way up to the podcasts of the present day. The Trove team couldn't pass up the chance to link the two systems.

Bringing this data in required thinking beyond the edge – the ABC makes this data freely available but it's not in a library metadata standard. The Trove team had never worked with a data set so large that wasn't in a library format, but with good metadata sharing principles embedded at the ABC end, Trove was able to capture and re-use the RN data.

Existing software at the NLA was adapted to harvest the rich metadata directly from the RN website. Past episodes of 84 separate programs were captured with Sitemaps. To stay up to date with new content, harvests were setup to check the RN RSS feeds on the same day a program is broadcast.

The first month saw 200,000 records harvested from RN and made available in Trove, growing to comprise more than 5% of the music, sound and video content available in Trove. This online, freely and immediately available Australian content found instant popularity with users – referrals to the ABC website climbed dramatically, tripling in that first month.

Discovery of RN content in Trove provides users with uninterrupted current affairs coverage and directs them to valuable resources on the RN website. Transforming to Dublin Core allows quantitative data analysis of these records, both as a set and in a broader context, which was not previously possible. This work has also opened the door for Trove to work with a greater range of institutions and their data – collections large and small – and underscores the importance of making structured metadata available, no matter the standard or format it's in.

This paper looks at the challenges involved in making big data accessible. How could we take the hundreds of thousands of program descriptions from the RN website and make them available to Trove users in a meaningful way – so they can discover the one little record in that big data set that is of relevance to them? How do we help digital historians find the answers, before they know what the question is? There are many more collections like that of RN – trusted, completely online and highly valued. This is one example of thinking beyond the edge of our system and the huge benefits it brought.

The NLA Harvester, OAI-PMH and the growth in Trove's content partners

The story starts with a piece of software called the NLA Harvester that was developed in-house at the National Library of Australia between 2007 and 2008. Its purpose was to read records from repositories and place copies of those records into NLA discovery services. Over its life the NLA Harvester has added records to Australian Research Online (ARO), Trove and Libraries Australia.

To understand how the NLA Harvester works, first think of a cultural institution as a big water reservoir at the top of a mountain, filled with fish (or metadata records). A pipeline, the NLA Harvester, is built with one end connecting to that reservoir and the other end connecting to the big ocean that is Trove. Records can now flow out of the repository through the NLA Harvester and into Trove.

The first connections to the NLA Harvester used the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH). The underpinning principles of OAI-PMH shape the NLA Harvester's core functions:

- Repositories are queried with HTTP GET requests;
- XML records are accepted as input;
- Each record is handled individually;
- Records are altered by successive transformations, either with java regular expressions or XSLT stylesheets;
- Updated records and delete commands are output to discovery services;
- Repositories are queried on a regular schedule for changes.

In the 7 years since this software was developed, other record-sharing technologies have grown in popularity among cultural organisations. The custom API is becoming increasingly common, the XML Sitemap schema is used by search engines, RSS feeds persist, and so too do plain lists of hyperlinks (or Harvest Control Lists).

There are similarities between these technologies and OAI-PMH. They all use HTTP and provide XML data. Building on that foundation a series of small upgrades were made to the NLA Harvester over the years allowing it to undertake selective web harvesting guided by Sitemaps, RSS feeds and Harvest Control Lists.

Despite these newer record-sharing technologies and the upgraded capability of the NLA Harvester, direct contribution to Trove was still predominantly via OAI-PMH at the end of 2013. By June 2012 there were 172 separate collections being harvested into Trove, 149 of those used OAI-PMH. Those 149 OAI-PMH users were concentrated in a few sectors: universities; national, state and territory libraries; public libraries; a handful of museums; and online biographical services (See Figure 1).

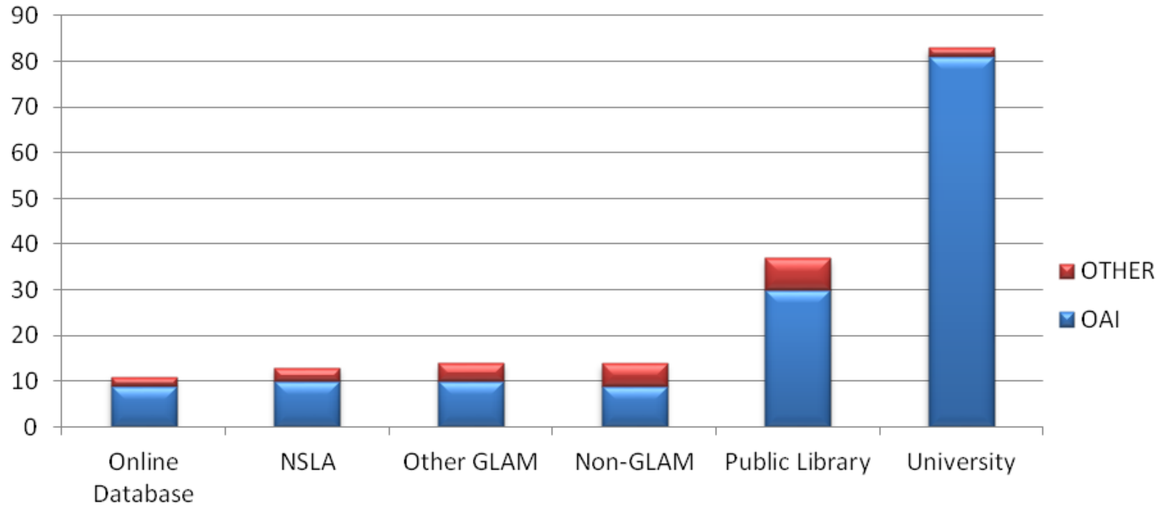


Figure 1
Collections contributed Pre-July 2012 via the NLA Harvester to Trove

Between July 2012 and June 2014 this started to change. New collections were increasingly coming from non-library organisations, who didn't use OAI-PMH (See Figure 2).

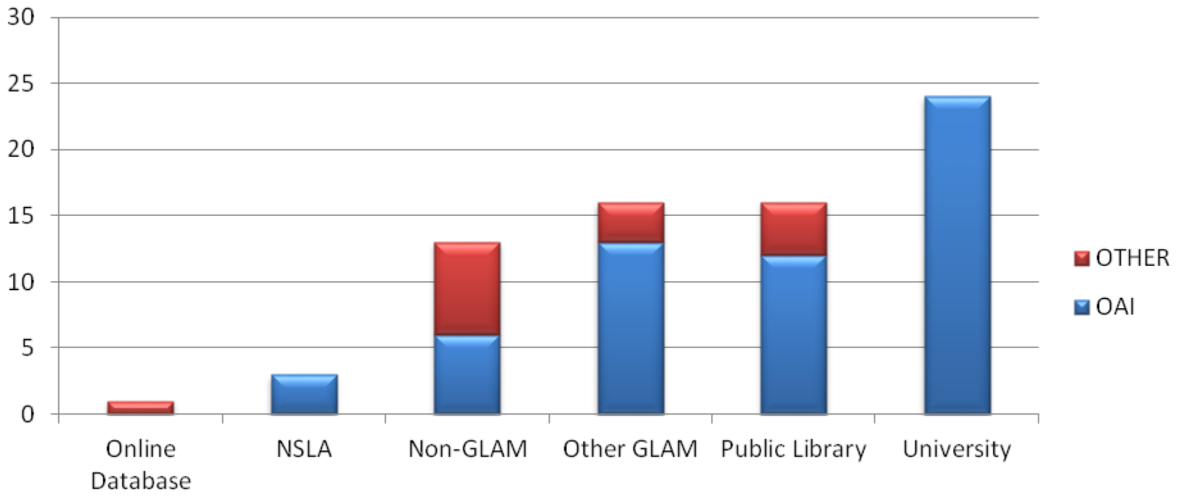


Figure 2
Collections contributed July 2012 - June 2014 via the NLA Harvester to Trove

Despite this shift, OAI-PMH remained the dominant method of contributing new collections. However we were aware that who *does* contribute is not always who *wants* to contribute.

An analysis of requests submitted to Trove's public contact form over the 2013/14 financial year revealed that lots of organisations want to get their content into Trove. In fact, 30 organisations registered their interest in becoming a content partner over that year and we didn't have a prior relationship with the majority. Of those 30 that approached us more than half were unable to make their collections available via OAI-PMH. It turned out that most of these organisations who could not provide OAI-PMH were also not libraries. (See Figure 3)

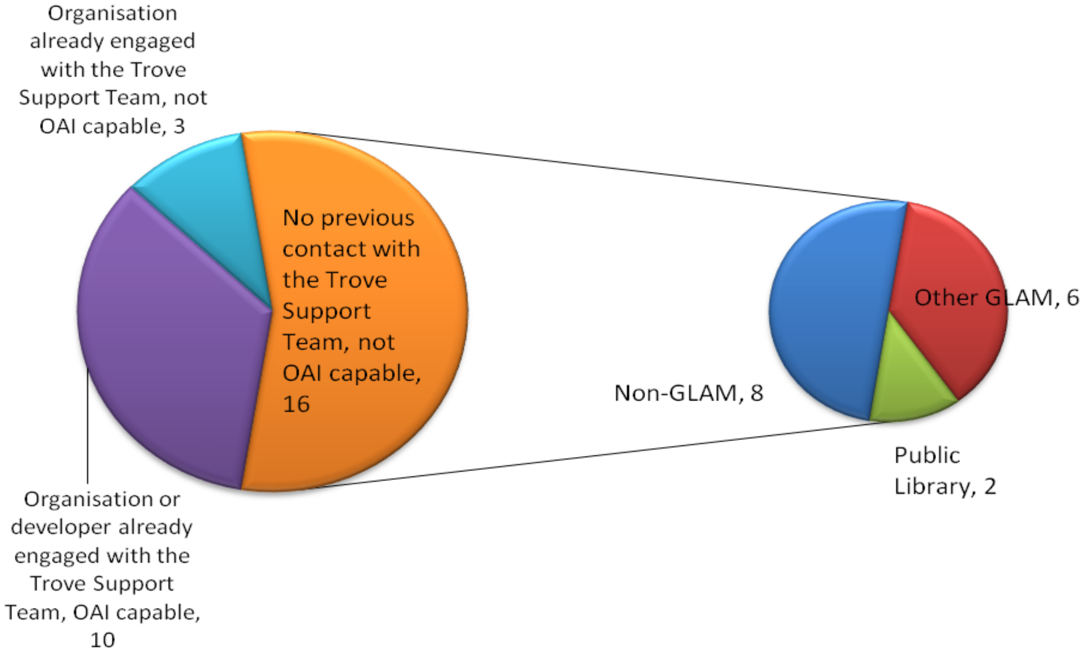


Figure 3
New potential collections for Trove June 2013 - July 2014

Worse, those non-library organisations who were so keen to join had the highest chance of never becoming a Trove content partner (See Figure 4).

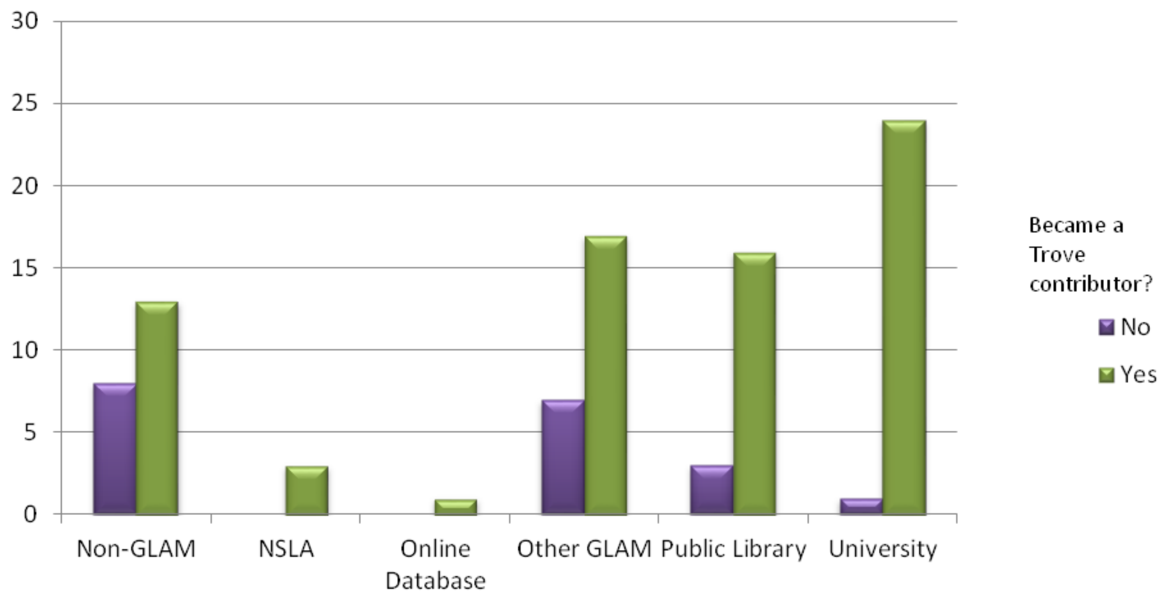


Figure 4
Potential contributors to Trove July 2012 - June 2014

Think back to that reservoir at the top of the mountain, with a pipe connected so that metadata fish can swim out into the big ocean that is Trove. Now we're seeing lots of different reservoirs popping up, wanting us to connect a Trove pipeline. They have a multitude of different coloured fish we want in Trove but it seems like they are surrounded by impenetrable rock. Our pipeline could reach them, we just lacked the specific connectors that would allow us to join the pipe to new and different types of reservoirs.

We were essentially faced with two challenges:

1. **Technical**
   In the past we built tools to work with technology that had been adopted by a single contributor, hoping it signalled an uptake by the wider community. We encouraged potential contributors to implement standards and embed metadata to work with the grain of the web. The benefits of this were greater than just being harvested by Trove, they would have improved crawling by the big search engines too.

   However potential content partners, especially smaller organisations, continue to indicate that making any change is an insurmountable hurdle, just as the "implement an OAI-PMH repository" request was last decade. If Trove can't accept their metadata as it is, then they usually don't join Trove.
2. **Perception**
   The good work of our predecessor systems in promoting standard protocols and prescriptive structured metadata saw organisations acquiring modules to specifically integrate with ARO or Picture Australia. That good work left an unintended legacy. Anecdotally we hear that organisations remember the significant challenges they faced in getting a pipeline connected. They don't approach Trove with new online collections that aren't OAI-PMH accessible, because of the belief that OAI-PMH is the only way Trove takes in metadata.

We want Trove to be a place to find unique Australian resources, no matter the technical abilities of the curating organisation. This led to a re-examination of the technology that the organisations have in common – a website. The websites we work with range from the front end of a specialised content management system to a standard WordPress installation. They all provide some form of metadata to give users context to an item. They may use structured meta tags, or they may not. Importantly, they all use HTTP to transfer webpages and rely on HTML, a form of XML, to display those pages in web browsers to users.

At the end of 2013 the challenge lay in working with those constants, the combination of HTTP and XML, and adapting the software we already had. Given the staff learning, investigation and setup time required this had to be a collection of considerable size and national significance to justify the investment.

An existing agreement had RN data slowly being added to Trove through an outdated, laborious and time consuming process that could not be completed without IT support. With a back catalogue of more than 200,000 segments yet to be captured, it was decided this collection would be a good place to start.

Harvesting Radio National

The ABC provides RSS feeds to keep up to date on recent episodes – perfect for our RSS feed reader input module. Past episodes aren't included in the RSS feeds though, so we needed something to capture that vast back catalogue.

We still wanted to use the NLA Harvester as a pipeline into Trove for the key workflows it provides, including:

- a simple scheduling interface;
- flexible record transformation options;
- robust output of records to Trove; and
- monitoring and management by the business area.

The RN website didn't meet the requirement of any existing NLA Harvester input module. We initially created mock RSS feeds at the NLA end, with links to every segment in a program's archive. This worked well for the smaller shows but proved too large for programs like AM with more than 40,000 segments in its archive.

We therefore took a step back and examined the website layout. It was a dependable hierarchy broken down into programs, then years, then segments (See Figure 5). We wanted to capture the bottom layer of that hierarchy, with each segment or episode to become a record in Trove.

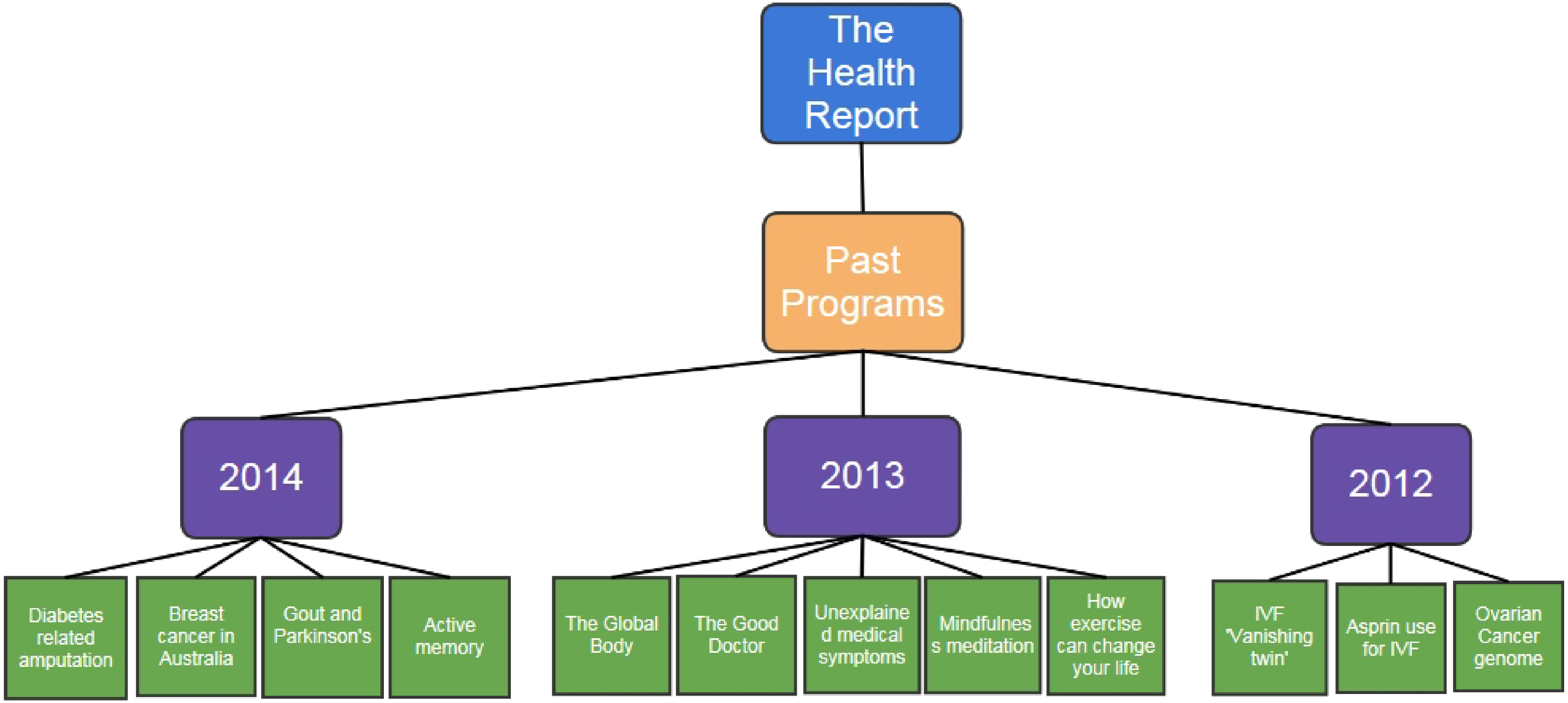


Figure 5

Logical Layout of a program on the ABC Radio National website

On analysing that website layout we realised this structure looked very familiar, just like a series of linked XML Sitemaps (See Figure 6). The only thing that was missing was the actual XML Sitemap files.

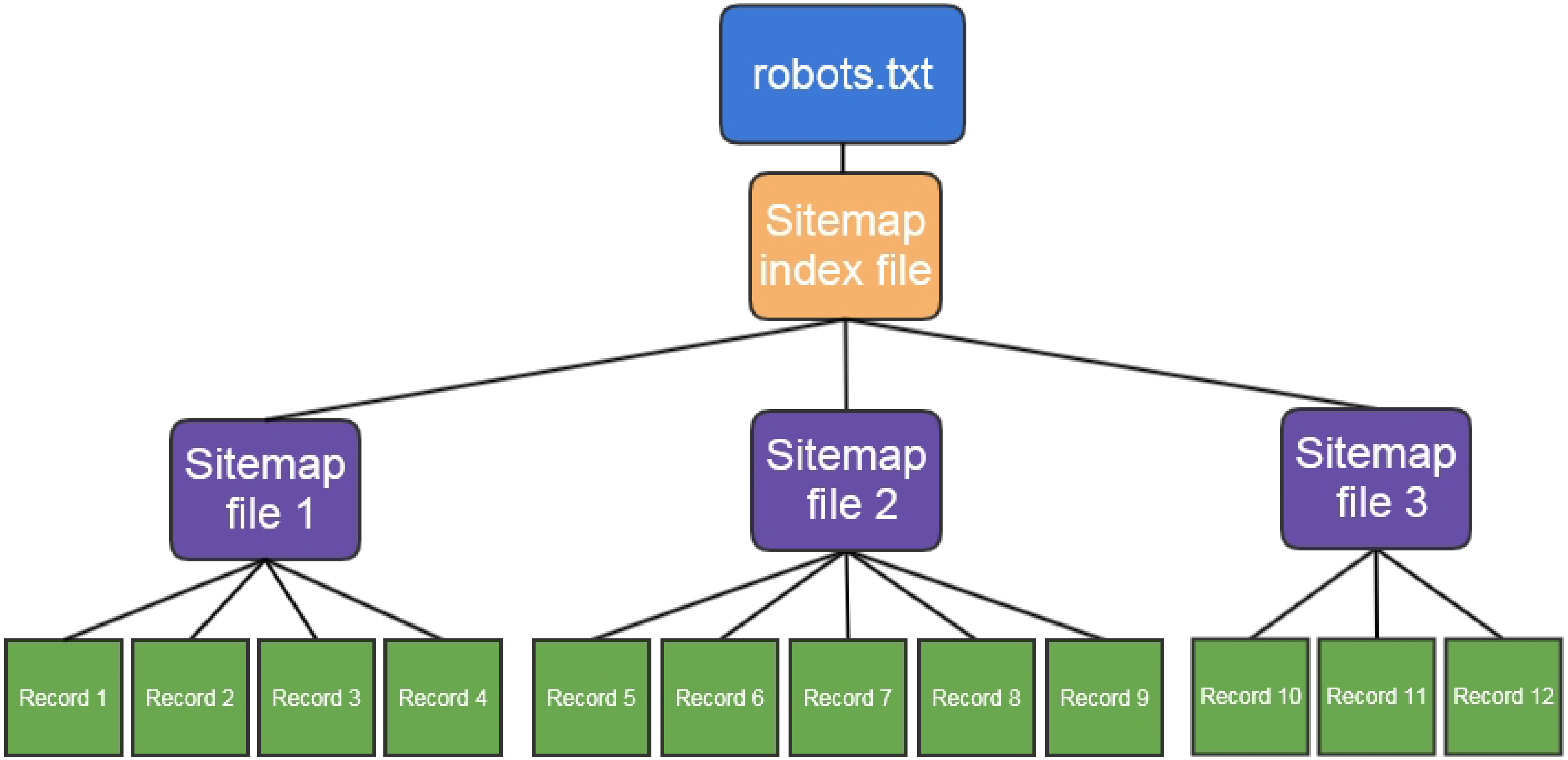


Figure 6
Generic XML Sitemap layout

We therefore prototyped a light, intermediary PHP script, a pre-processor to lie between the NLA Harvester and the RN website. This PHP script is called by the Harvester and in turn goes to the live RN site. It turns the list of years in the past program archive into a Sitemap index file and generates plain Sitemap files composed of links to each segment or episode broadcast in a single year.

Now we had Sitemaps in the structure the NLA Harvester required. The RN Sitemap script is effectively a small piece of pipe connector, an interface between the RN reservoir at the top of a mountain, and the opening of the NLA Harvester pipeline.

Once the harvester had captured the data from the RN website, it needed to be transformed into records that Trove could use. Specifically, we had taken a copy of the HTML page from the RN website and needed to convert it into a Dublin Core XML record to be loaded into Trove.

When we looked at a segment on the RN website we could see details like the title, date of broadcast and transcript. This information appeared throughout the HTML document and not as a structured XML metadata record that Trove could accept. To create a Dublin Core XML record suitable for loading into Trove we needed to capture the relevant information from the HTML document using the NLA Harvester's conversion tools.

As an example, here is some of the information from the segment *Black diggers* (http://www.abc.net.au/radionational/programs/bigideas/black-diggers---abc-symposium/5733810) on the program *Big Ideas*:

- Title: Black diggers
- Presenter: Paul Barclay
- Broadcast date: Thursday 25 September 2014
- Guests: Uncle Dave Williams, Lee-Ann Buckskin, Professor Lisa Jackson-Pulver, Wesley Enoch, Katrina Sedgwick

These pieces of information are useful for finding a known item, where the user is searching for a specific broadcast that they already know exists. For the user searching more broadly by subjects or keywords, like "Indigenous" or "World War One", they're not going to find this record. We need to capture more information, so that broader searches will find this record in Trove. Luckily the ABC has included a large number of <meta> tags for each segment. These tags aren't viewable to an ordinary user with their web browser but include additional helpful information – subjects, a brief description of the segment and more – that facilitate better discovery. To create a record for Trove, we capture data from both the elements displayed to users and this hidden data from <meta> tags.

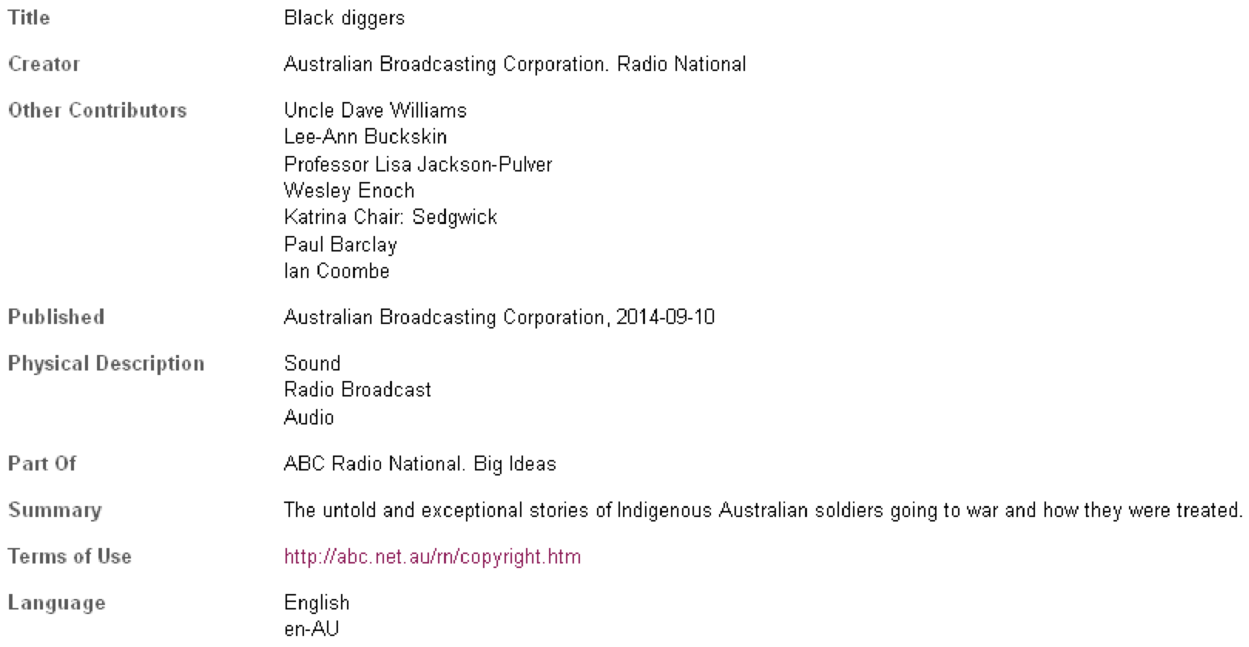
The first step in creating any record is to examine the data that we are receiving, identify the elements we want and create a map to Dublin Core. Some of the fields that we wanted to capture from the RN website were:

- Title – Title of the segment, captured from <meta> tags;
- Contributor – names of guests, captured from the page content; and
- Subject – subjects and keywords assigned by the ABC, captured from <meta> tags

Capture of these fields would allow us to create a record that a user could discover in Trove, and assess its suitability. To make it even easier for users to find the records in Trove we also take a copy of the full transcript (where available) and index this text for more relevant search results. This text is not made visible in Trove.

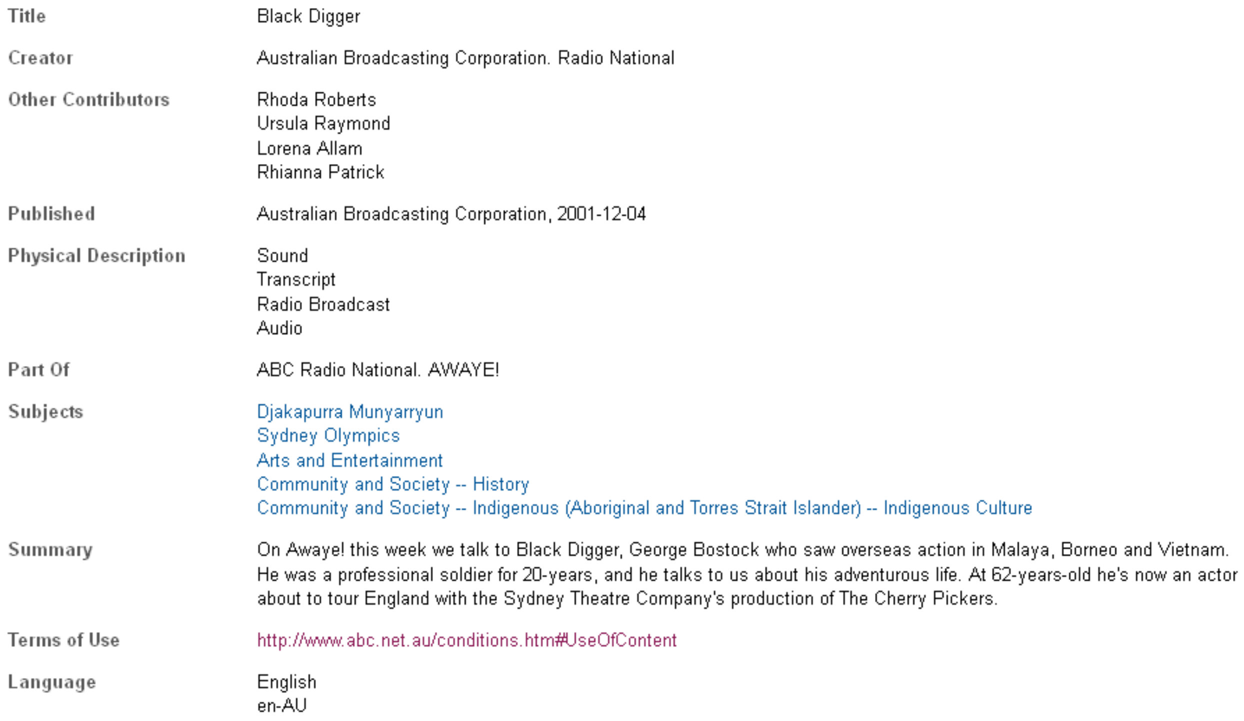
Prior to working on the RN content, records captured from HTML pages typically relied on data in <meta> tags, where organisations had been instructed on which tags to use for best results in Trove and its predecessors. When looking at the RN content we had to think outside of this approach, and investigate ways to transform additional useful data beyond that contained in <meta> tags. We started working with the structural elements of the HTML page to identify and capture the relevant data.

As all RN shows are stored in a common content management system (CMS), the layout of the pages is uniform. Therefore a XSLT stylesheet written to capture data from one show could be used to capture the data from any show stored in the same CMS. This allowed for a single approach to create homogenised records from different shows, as shown in figures 7 and 8.



Title	Black diggers
Creator	Australian Broadcasting Corporation. Radio National
Other Contributors	Uncle Dave Williams Lee-Ann Buckskin Professor Lisa Jackson-Pulver Wesley Enoch Katrina Chair: Sedgwick Paul Barclay Ian Coombe
Published	Australian Broadcasting Corporation, 2014-09-10
Physical Description	Sound Radio Broadcast Audio
Part Of	ABC Radio National. Big Ideas
Summary	The untold and exceptional stories of Indigenous Australian soldiers going to war and how they were treated.
Terms of Use	http://abc.net.au/rn/copyright.htm
Language	English en-AU

Figure 7
Record in Trove for the item ‘Black Diggers’ as broadcast on the show ‘Big Ideas’
http://trove.nla.gov.au/version/209294503



Title	Black Digger
Creator	Australian Broadcasting Corporation. Radio National
Other Contributors	Rhoda Roberts Ursula Raymond Lorena Allam Rhianna Patrick
Published	Australian Broadcasting Corporation, 2001-12-04
Physical Description	Sound Transcript Radio Broadcast Audio
Part Of	ABC Radio National. AWAYE!
Subjects	Djakapurra Munyarryun Sydney Olympics Arts and Entertainment Community and Society -- History Community and Society -- Indigenous (Aboriginal and Torres Strait Islander) -- Indigenous Culture
Summary	On Awaye! this week we talk to Black Digger, George Bostock who saw overseas action in Malaya, Borneo and Vietnam. He was a professional soldier for 20-years, and he talks to us about his adventurous life. At 62-years-old he's now an actor about to tour England with the Sydney Theatre Company's production of The Cherry Pickers.
Terms of Use	http://www.abc.net.au/conditions.htm#UseOfContent
Language	English en-AU

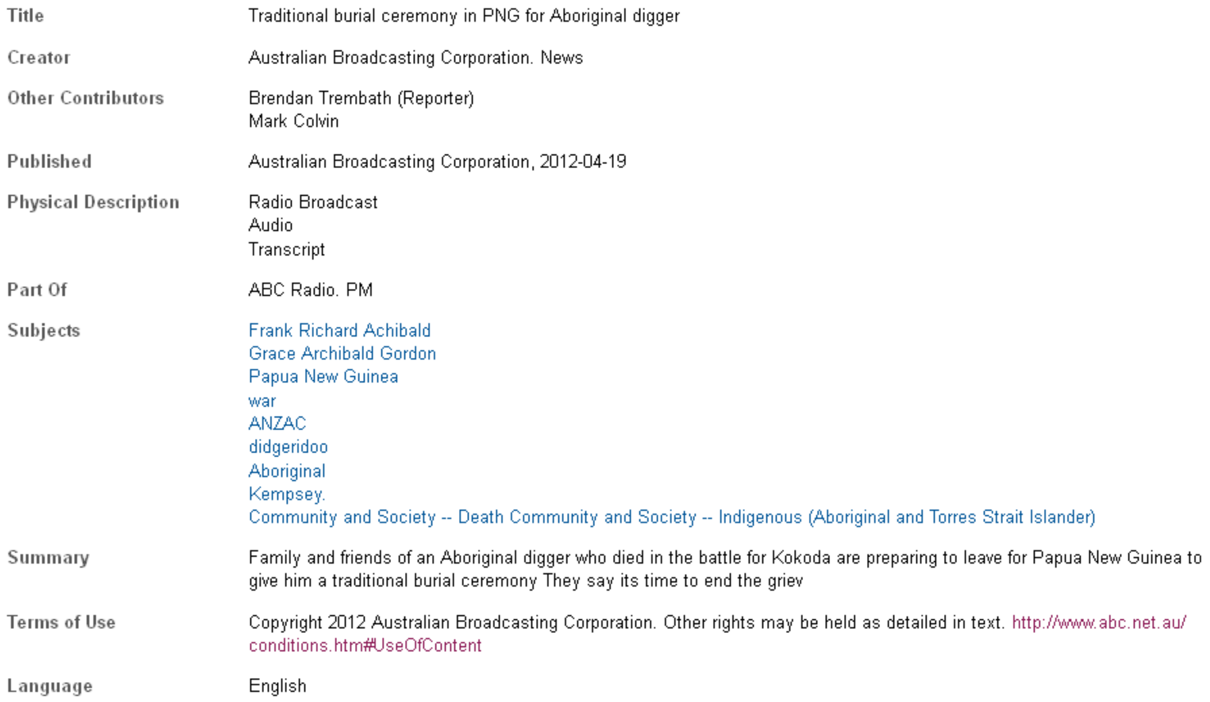
Figure 8
Record in Trove for the item 'Black Digger' as broadcast on the show 'AWAYE!'
http://trove.nla.gov.au/version/204814184

When harvesting these records we discovered, through a segment of *Encounter*, that there was a limit on the size of records that the NLA Harvester could process. Luckily very few of the segments in the Radio National collection approach the limit, but to avoid this issue we have needed to limit the length of the transcripts for these segments. This has had no noticeable impact on discovery of these segments with keyword searches.

This approach worked well for the RN shows, however when we started looking at current affairs shows (AM, PM, The World Today, Correspondents Report) we found a number of other challenges. The first was that the pages for these shows are not stored in the RN Content Management System. The page layouts were quite different, and therefore needed a separate set of transformations. Additionally, although most segments from 1999 onwards are available online, the page layout varies significantly over time. This complicated the process of creating consistent and accurate records for Trove users.

Although a lot of data for the current affairs shows was included in <meta> tags, some important information was only available within the transcript. To create accurate records in Trove, we therefore needed to think outside our normal processes and develop a process that would allow us to capture structured data from the transcript.

This process allowed us to create records for current affairs shows that contained the same elements as the records for other RN shows (See Figure 9).



Title	Traditional burial ceremony in PNG for Aboriginal digger
Creator	Australian Broadcasting Corporation. News
Other Contributors	Brendan Trembath (Reporter) Mark Colvin
Published	Australian Broadcasting Corporation, 2012-04-19
Physical Description	Radio Broadcast Audio Transcript
Part Of	ABC Radio. PM
Subjects	Frank Richard Achibald Grace Archibald Gordon Papua New Guinea war ANZAC didgeridoo Aboriginal Kempsey. Community and Society -- Death Community and Society -- Indigenous (Aboriginal and Torres Strait Islander)
Summary	Family and friends of an Aboriginal digger who died in the battle for Kokoda are preparing to leave for Papua New Guinea to give him a traditional burial ceremony They say its time to end the griev
Terms of Use	Copyright 2012 Australian Broadcasting Corporation. Other rights may be held as detailed in text. http://www.abc.net.au/conditions.htm#UseOfContent
Language	English

Figure 9

Record in Trove for the item 'Traditional burial ceremony in PNG for Aboriginal digger' as broadcast on the current affairs program 'PM!' http://trove.nla.gov.au/version/204848635

Why?

The purpose of Trove is to connect people with resources, whether this is through the Trove user interface or through applications developed using the Trove API. Although some specialised content is delivered through Trove, the focus is on discovery, not delivery. Content that we include in Trove from our content partners is included to extend discovery of these resources outside the dedicated audiences of that organisation. We only receive metadata for items from content partners, and push users to that organisation where they can listen to, view or read the entire online item.

Contribution of records to Trove provides content partners with a higher return on the investment that they put in to collecting, maintaining, cataloguing and delivering their collection by driving more users to visit their website. This return on investment is possible because Trove is a resource accessed by a large variety of people. By including resources from small organisations, such as the Queensland Police Museum, alongside the resources held by the National and State Libraries, the Powerhouse Museum, the Australian War Memorial and Museum Victoria, these resources are discovered in the broader context of Australian collections. Trove is contributor neutral, so users will always find the most relevant results for their search.

By leading users to quality resources from Australia's collecting institutions, Trove has built a reputation as a place to start a search for high-quality information. The ABC is held in similar regard by many Australians, so these records are a natural fit for Trove.

Inclusion of RN records in Trove means that new users are more likely to discover this collection. Although these users start in Trove, they are likely to further explore the content partner's site for more resources on the same topic. This holds true for any collection we include in Trove and is described by Michael Neubert in his blog, *From Wheels to Bikes* (http://wheelbike.blogspot.com.au/2012/02/starting-search-for-bikes-in-trove.html), where his discovery of photographs in Trove lead him to Gallica (from the Bibliothèque nationale de France). If these items had not been included in Trove, Michael would have missed a significant source for his research. Inclusion of RN records provides this kind of serendipitous discovery of resources, as although users may think to look at the AM, PM or World Today subsites for current affairs information, they may not think to look to the RN website for information about schizophrenia, http://trove.nla.gov.au/music/result?q=nuc%3AABC\%3ARN+schizophrenia; Aboriginal diggers, http://trove.nla.gov.au/music/result?q=nuc%3AABC\%3ARN+Aboriginal+diggers; or the political legacy of Dr Seuss, http://trove.nla.gov.au/music/result?q=nuc%3AABC\%3ARN+%22dr+seuss%22+political.

When a new collection is made available in Trove we see an immediate impact, with users finding and clicking through to this content almost immediately. This can be seen when looking at the click-through stats for the ABC from Trove. In January 2014 we see a significant, and sustained, increase in the number of click-throughs to the ABC from Trove, which coincides with this content being made available (See Figure 10).

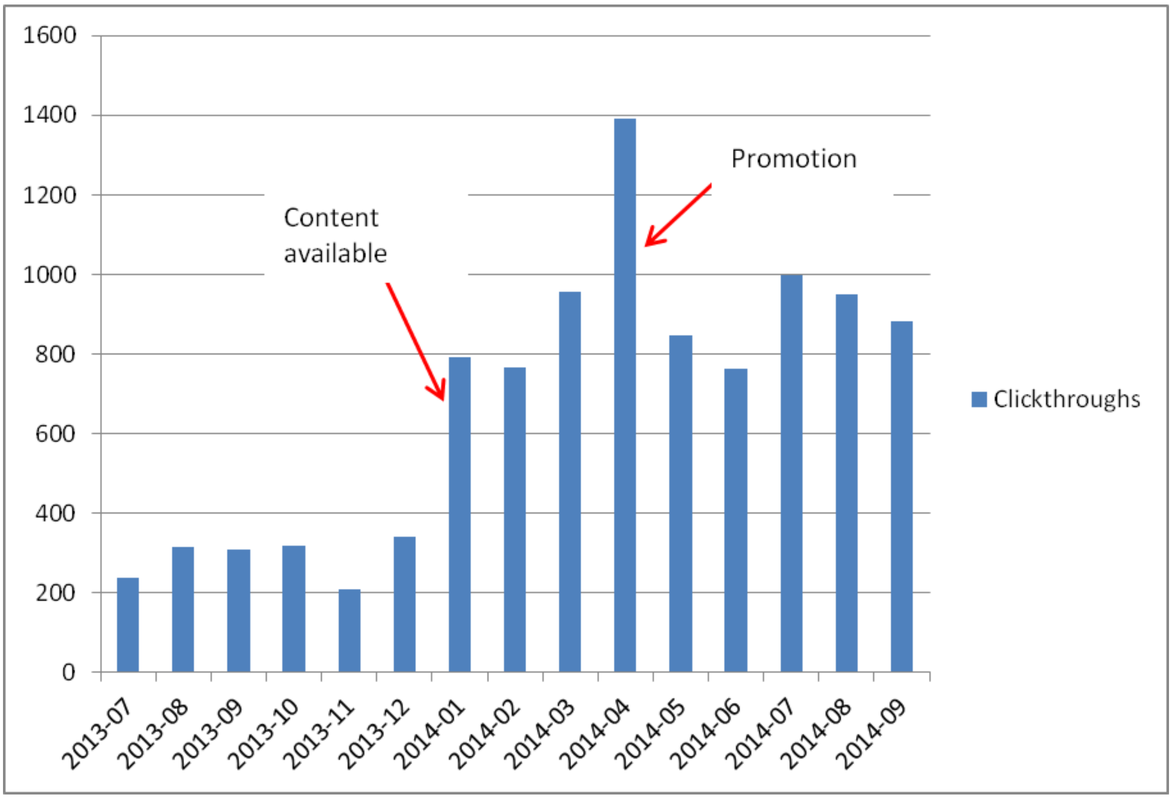


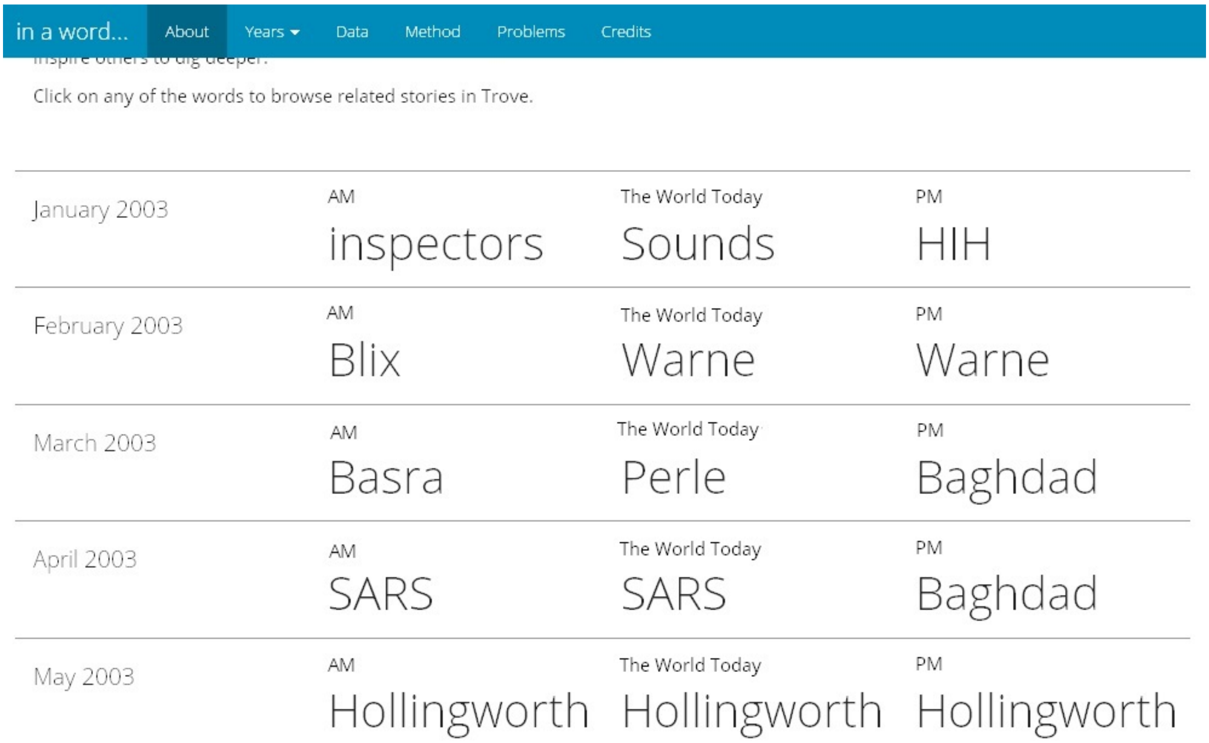
Figure 10
ABC Click throughs July 2013 – June 2014

We do see a peak in click-throughs in April, when we were heavily promoting this content; however the increase from January has been sustained.

What is the benefit to Trove of including the RN resources? In making the decision to capture these resources a number of factors were taken into account. The first of these is simply coverage. By including the current affairs shows (AM, PM, The World Today) in Trove we extend current affairs coverage through to the current day. We also make the data from RN available for reuse through the Trove API. Another major factor in the decision to do the required work is that the inclusion of RN content aids in diversifying the media types available through Trove. In fact RN records now account for almost 6% of all content in the Music, Sound, Video zone of Trove with 208,000 records.

By making the RN records available through the Trove API, they are opened up for additional uses and analysis not previously possible. One possible use is for libraries to include records from RN in their catalogue, allowing users to find these resources alongside other items held in local collections. Digital historians are also able to use these resources for further analysis.

By having a standardised set of RN records available through the Trove API, in particular the records for the current affairs shows, digital historians are now able to easily perform analysis on the content of the records. One such example is Tim Sherratt's *In a Word'* (http://inaword.dhistory.org/, Figure 11), a textual analysis of AM, PM and the World Today that identifies distinctive words to give an overview of the main themes of the current affairs programs for each month from 2003 to 2013



inspire others to dig deeper.

Click on any of the words to browse related stories in Trove.

January 2003	AM inspectors	The World Today Sounds	PM HIH
February 2003	AM Blix	The World Today Warne	PM Warne
March 2003	AM Basra	The World Today Perle	PM Baghdad
April 2003	AM SARS	The World Today SARS	PM Baghdad
May 2003	AM Hollingworth	The World Today Hollingworth	PM Hollingworth

Figure 11
Tim Sherratt's 'In a Word' analysis of the most common word for each month

Lessons Learned

Through the work that we've done with the ABC to bring the RN content into Trove we have learned a number of lessons on how to capture available metadata. One of the most important is the value of structural metadata that was available in all the RN pages. Both the metadata in a traditional sense, and exploring beyond that to the useful structural elements - for example having named <div> elements - contained within the pages allowed us to target specific locations on the page and capture the data that we wanted to include in Trove records. This technique can be applied to many more websites built on top of content management systems.

The use of structured metadata on the ABC websites encouraged the Trove team to re-examine data sources previously considered unsuitable for Trove. Without the work that was completed to bring the RN content in a number of other data sources, such as the Australian Government Solicitor's Legal Opinions website, the Australian Parliamentary Library's Press Releases database and the AusStage events database, would not have been considered as suitable sources and a lot of valuable data may not have been included.

Another lesson we learned is that even when records are captured with website harvesting rather than structured data retrieval, a CMS with organised data is still an important foundation for record sharing. The reason for this is that with a CMS each element of the page contains the same type of content. This allows us to reliably look in a single place for a piece of data instead of specifying many variations, minimising the risks associated with harvesting data not solely presented as a structured metadata record. Changes to the website are likely to be aesthetic only, and not significantly impact on the data contained.

There are maintenance costs associated with the capture of data from the ABC, we need to monitor the harvests and check to make sure that records are being correctly captured. This ensures that the harvests are working as we expect and that any changes made to the pages are accounted for, allowing us to continue to capture all relevant data. These maintenance costs apply to other forms of harvesting where systems change over time, so the harvests for RN content are not unique in this respect.

Another lesson that we learned in the development of processes to capture the RN data is that we can take our existing systems beyond the limits that were originally conceived. In this specific case, we made use of the Sitemap and RSS harvesting functionality to provide the harvester with a harvesting target, even though the Sitemap was not available on the original site. The next step was to use the transformation functionality of an XSLT stylesheet to transform the content, even though this functionality was originally designed for an XML structured metadata record rather than a HTML page.

Finally, we've moved away from the earlier style of development where we poured IT resources into creating modules that served a single new content partner, and saw little re-use by later contributors. Instead we've successfully prototyped a tool for one collection, ABC Radio National, then gone through iterations to make the script more generic. This script can now serve many more organisations and sectors. We've been able to do real-world testing and develop much more precise requirements than we otherwise could have at the outset. Now, we can use those requirements to guide further long-term development.

The Future

With this script we can now look at saying 'yes' to more of those potential content partners who cannot implement OAI-PMH. In particular, we want to re-examine those past potential contributors who couldn't meet the technical requirements and ask the question – "Could we use the script to capture this website?” Inclusion in Trove can be the start of a metadata sharing experience for an organisation, demonstrating the value of exposing collections to a broader audience and drawing more users back to their websites. We would like to provide results that drive further organisational changes and result in better exposed and maintained metadata.

We want to encourage more libraries and scholars to re-use this dataset. Just as Trove takes in data from the ABC, so too can anyone take that data from Trove. The records are now in a library-standard format, Dublin Core, and available through the Trove API along with all other Trove records. They can also be accessed as a stand-alone collection. One day soon we’d like to see Library Catalogues using that data, displaying ABC records alongside other search results or as a freely available alternative to subscription databases. We’d also like to see more scholarship focused on analysis of large datasets like this one, leveraging off the work done to normalise the records and make them easily accessible.

Software capable of harvesting from a wider range of organisations is a necessity. Like many other organisations Trove has to work within its budget, servicing our audience with existing resources. Projects like this one help us to do that – using existing software, existing functionality and existing robust tools to meet new requirements, work with new content systems, bring in new contributors and deliver up new data in Trove. We will continue to experiment with smaller projects, helping good ideas to flourish while poor ones fail quickly without a significant outlay of time or money. Small projects help us to better plan where we expend our limited IT resources in future upgrades, focusing on the tools that have already proved they will offer a high return on investment.

Most importantly, we are trying to take the blinkers off and think beyond conventional data types and content partners. Bringing in the ABC Radio National data has underscored that high quality, free, Australian resources from trusted institutions are valued by the Trove users who have a voracious appetite for new data. We now plan to seek out other strategic partnerships and look at all new datasets, collections and institutions with fresh eyes, with a mindset not of “how do they need to change to fit Trove” but rather “how can we use Trove’s tools to work with them”.