Extending the scope of Trove: Addition of e-resources subscribed to by Australian libraries

(Final Copy for publication in Nov/Dec 2011 issue of D-Lib) http://www.dlib.org

Rose Holley, Trove Manager, National Library of Australia rholley@nla.gov.au
2 November 2011

Word count: 4600
Tags: Licensed e-resources, subscription e-resources, Trove, National Library of Australia.

Abstract:
Trove is the national discovery service for Australia managed by the National Library of Australia. It was released in December 2009. It contains metadata for millions of freely accessible items, from more than 1,000 contributing institutions. The focus is on Australia and Australians. In 2011 it was developed further to include selected sets of e-resources subscribed to by Australian libraries. Trove v4.0 was released in May 2011 after 120 million subscription e-resources were successfully included. This took the Trove content total to almost 240 million items. This article describes why and how the work was undertaken, what was achieved, the issues and future plans for development.

1. BACKGROUND

Trove (http://trove.nla.gov.au)¹, the national discovery service for Australia was released by the National Library of Australia in December 2009. At that time it contained metadata for 80 million freely accessible items, including those harvested from 1,000 contributing Australian institutions. The focus was on Australia and Australians. Trove demonstrated innovation in resource discovery and access².

In 2010 it was decided to extend the scope of Trove to include selected sets of e-resources subscribed to by Australian libraries. The work to implement this was undertaken over six months from November 2010 – May 2011, in partnership with the National State and Territory Libraries of Australasia (NSLA) Reimagining Libraries Open Borders Project³ and two e-resource vendors, Gale Cengage and RMIT. A primary goal of the development work was to see if a theoretical model for national access of subscription resources translated into a practical workable model.

The work was successful and in May 2011 version 4.0 of Trove was released which contained 120 million subscription e-resources for Australians. This took the Trove content total to almost 240 million items. Extending the scope of Trove to include e-resources subscribed to by Australian libraries was part of the National Library’s strategic planning⁴. Australian users of Trove are now able to access subscription e-resources within Trove when they are a member of an Australian library that has subscribed to a product included in Trove.

This article describes why and how the work was undertaken, what was achieved, the issues and future plans for development.
2. WHY EXTEND THE SCOPE OF TROVE?

The two main reasons to extend the scope of Trove to include subscription e-resources were to:

- give Australian library users improved access to e-resources, especially those e-resources subscribed to by NSLA member libraries and Australian public libraries; and
- ensure that subscription e-resources are used to the maximum extent possible, by supporting a user-centric approach to their discovery and access.

This is reflected in the Library’s strategic objectives for 2009 – 2011:

1. We will collect and make accessible the record of Australian life...We will explore new models for creating and sharing information and for collecting materials.
2. We will meet our users’ needs for rapid and easy access to our collections and other information resources.
3. We will collaborate with a variety of other institutions to improve the delivery of information resources to the Australian public.

3. PREPARATION

A survey of NSLA member libraries was carried out in April 2009 by the Open Borders Project Group. It sought information on e-resource management solutions used by NSLA member libraries including: the current status of OpenURL link resolver services, barriers to easy access to articles contained in licensed e-resources, and authentication services. It asked libraries to identify the benefits of collaborating to provide national access to e-resources via Trove. Key benefits identified were to increase the visibility and use of subscription resources nationally; to help public library users access e-resources; to streamline access for readers who use multiple NSLA member libraries; and to share the work required for authentication and access. There was clear support for a more collaborative approach.

The survey results identified why librarians thought users may not be using subscription e-resources from their libraries. These included reasons such as lack of knowledge of the existence of e-resources in libraries by most users; limited or no promotion of subscription e-resources; the need to register, login and remember multiple passwords; and the lack of access if the user was off-site.

It appeared there were no solutions in place that could be easily extended to support federated authentication. Initial ideas the National Library explored were to establish a national proxy service, requiring screen scraping of logins at participating library sites and the creation of a shared Australian OpenURL Resolver infrastructure; and the development of a Knowledge Base for e-resources core to NSLA members and Electronic Resources Australia (ERA), that might meet the needs of NSLA and public libraries who do not already have an OpenURL resolver. After investigation and discussion it was decided not to take this approach because of the resources required to build and maintain the Knowledge Base. A less costly approach was sought. By this time it was clear that Trove would become part of the solution.

The agreed and less costly approach was to establish an “access and linking framework”
based on Trove. In this framework, the authentication would be pushed back either to the user’s library (via EzProxy or similar services) or to the vendor (using IP authentication or user barcode), rather than maintaining a national proxy server and OpenURL resolver. A decision was made not to use libraries’ OpenURL resolvers, but instead to link to the article on the vendor’s database because we know it is there. Some OpenURL resolvers are not well maintained, and the OpenURL isn’t a great standard in practice. For example, variations in article title between the vendor’s metadata and the vendor’s index and OpenURL knowledge base often mean an article will not be found with an OpenURL search.

Article metadata would continue to be maintained by e-resource vendors, and the National Library would develop a set of partnerships with these vendors, rather than attempt to maintain a shared national Knowledge Base. Two vendors (Cengage Gale and RMIT Publishing) agreed to work with the Library under this framework to expose their e-resource content in Trove.

Under this framework:

- article-level metadata, including the vendor’s article URL and full-text articles for indexing, where available, would be provided by the vendor to the National Library for inclusion in Trove and stored on the Library’s servers;
- vendors would also supply data about which articles are in which products and which products are licensed by which libraries;
- users of Trove would be encouraged to register with Trove and within their user profile provide information about which libraries they are affiliated with;
- Trove would index the article-level metadata (and full text where available) and would use the subscription and affiliation data to give “Available online” status to those articles which the user is entitled to click through to and read;
- Trove would facilitate the process of authenticating the user by either routing the user through their affiliated library’s proxy server, or telling the vendor the user’s library affiliation; and
- Trove would refer the user to the vendor’s site, where any remaining authentication and access to the full text would be managed.

Instead of developing a new database to store libraries’ IP and authentication details, which would have been significant work, it was decided to develop the existing Australian Libraries Gateway (ALG)\(^7\) to support the authentication. Trove already makes use of the ALG to retrieve information about libraries for display in Trove. The ALG had already been enhanced to support features needed in Trove to improve ‘getting’ options\(^8\), such as clear, short names for libraries for display in Trove’s brief results sets. Another advantage of using the ALG is that libraries are able to update their own details. For the development work however, the National Library intended to enter details on behalf of libraries.

### 4. IMPLEMENTATION PLAN FOR ADDITION OF SUBSCRIPTION E-RESOURCES INTO TROVE

Implementation tasks included enhancing the Australian Libraries Gateway and Trove, developing Trove system architecture and infrastructure, obtaining subscription content, testing the implementation and marketing the outcomes. The tasks are summarised in the table below.
1. **Enhance Australian Libraries Gateway (user authentication)**

Add new fields to the ALG under “Trove configuration” (customer IDs, proxy server information and IP address ranges) and then discover and add the following information for Australian libraries:

- Proxy server addresses and configuration details. This enables Trove to direct users to their library’s proxy server, correctly passing the access URL for the selected article.
- Subscribing library’s IP addresses to enable Trove to determine if a user is virtually or physically “onsite” at their library.
- Vendor customer numbers. These identify the library to the vendor.

2. **Enhance Trove**

- Create a new zone for Journal articles (because 120 million new articles will swamp content in the existing books and journals zone, making books harder to find).
- Migrate existing journal content in Trove to the new zone.
- Map the various proprietary formats to a common data format.
- Redesign the Trove home page and underlying pages to accommodate the new zone.
- Revise the search results and full view screen to accommodate the new data.
- List libraries that have copies of the journal an article belongs to.
- Provide a link to the National Library’s “Copies Direct” service for articles likely to be held by the National Library.
- Create two new facets in the “Journals, articles and datasets” zone to help users limit their search:
  - Audience (only applies to articles from Gale products, e.g. “Professional”)
  - Database (e.g. “Australian Public Affairs Full Text”)
- Display the ‘short library name’ on the results screen and the ‘get’ screen, so users can choose the library subscription they have access to.
- Match IP address of user against information in the ALG to determine if a user is ‘onsite’ at a subscribing institution.
- Pass article information to the institution’s proxy server so that users can log in and immediately access the article, or pass the article information and user affiliation to the vendor’s logon page for vendor to authenticate the user.
- Implement full text indexing where full text is available.
- Process subscription and data updates as received.

3. **Develop Trove system architecture and infrastructure**

- Ingest and store the full-text of articles on Trove servers.
- Re-organise the Trove Lucene search indices.
- Create a new Lucene index of 400GB to store the article search information.
- Index 120 million newly ingested articles.

4. **Obtain subscription content**

- Obtain lists from RMIT Publishing and Gale Cengage of their Australian customers and their customer codes.
- Create mappings from library (NUC) codes to the vendor customer codes.
- Obtain from RMIT Publishing and Gale Cengage the metadata and full-text of articles in an agreed list of products.
- Obtain subscription and data updates from vendors on a regular basis.
- Vendors may wish to contact their customers and let them know of the work being undertaken in Trove on licensed e-resources.
5. **Testing of implementation**

Ask selected libraries to test their e-resource subscriptions in Trove.

6. **Marketing outcomes**

- Inform libraries that their e-resource subscriptions are now available in Trove.
- Inform Trove users that Trove now includes subscription e-resources.

The following diagrams show the current Trove infrastructure and how users access articles. **Figure 1. Trove Infrastructure**
5. ACHIEVEMENTS

On 12 May 2011, after six months work by a four-person development team, Trove 4.0 was released. This expanded Trove by adding article-level metadata from five Gale databases and nine Informit databases, a total of about 120 million articles. This initial load included Gale’s Academic OneFile, General OneFile, Literature Resource Center, Health and Wellness Resource Center and Expanded Academic ASAP. Also Informit’s Australian Public Affairs Full Text and APAIS records, Australian Medical Index, Humanities and Social Sciences Collection, Health Collection, Engineering Collection, Business Collection, New Zealand Collection, Meanjin Backfiles and Media International Australia Backfiles were added.

These e-resource articles are now accessible online in Trove to users affiliated with a subscribing library. If users are onsite at their library and their IP address has been registered in the ALG, or they have indicated which libraries they are affiliated with (by logging in and...
setting up “my libraries”), Trove will know which articles they can access and display this information. Trove will:

- Change the online status of these articles from “access conditions apply” to “freely available”, showing a bright green “View online” rectangle against these articles on the results screen.

- When displaying the article details, show the user’s libraries and a prominent corresponding link to access the article at the top of the “online” tab.

If users are not logged in to Trove when viewing the article details in Trove results, they are able to access the full-text article by clicking the link which says ‘show me all libraries that have a subscription to this article’ and then choosing a library they are a member of from the list. They will then be requested to login with their member library details to access the article.

Users can be successfully authenticated by Gale and Informit in various ways:

- Automatically (by IP address), when onsite at their library;
- Through their library’s proxy server, which they must log into; or
- Using a barcode or password entered on the Gale or Informit site.

**Figure 3.** Screenshot of Trove showing what a member of the National Library of Australia would see for a subscription e-resource. The National Library access link appears above the link button to see all other libraries with subscriptions.
Where possible, when showing the article details, Trove also displays library holdings of the journal the article appears in. Articles appearing in a publication held by the National Library can be requested through the Library’s Copies Direct service (shown on the “buy” tab when viewing the article details in Trove). If a user is not affiliated with a library that has an Informit subscription, then individual Informit articles from most databases are available to purchase on demand.

There are 1,000 Australian libraries contributing to Trove. Nearly half of these subscribe to Gale or RMIT products. The vendors provided the Trove team with their Australian customer lists, which the Trove team then matched with libraries in the ALG. These customer lists included public, state and territory, school, tertiary, university, government and private industry libraries. Over 450 libraries had their authentication details added to Trove and are now able to access their subscription e-resources via Trove.

6. ISSUES AND THEIR RESOLUTION

New Trove zone
Before the development work full text articles were contained in one of in two zones: ‘Books, journals, magazines and articles’ (for harvested data), and ‘Digitised newspapers and more’ (for Australian digitised historic newspapers and the Australian Women’s Weekly: outputs of the Australian Newspapers Digitisation Program). There were already identified issues by having similar content in two different zones. Early on it was realised that the volume of subscription e-resources was potentially larger than existing Trove content, and it was essential to create a new zone (and Lucene index) for them. If this had not been done the articles would have swamped the books zone, making books harder to find and also negatively affecting system performance. The impact of creating a new zone was considerable on system work, users, and marketing. The team discussed the name of the new zone and what to include or not include in it. It was not feasible to include the 50 million newspaper ‘articles’ in this new zone. After much discussion it was agreed to migrate all existing Trove articles into a new article zone (with the exception of newspaper articles), whether or not they were subscription or freely available. The new zone was called ‘Journals, articles and data sets’.

Finding Proxy server addresses
The Trove team had to establish almost 100 proxy server addresses to enter into the ALG. This was mostly done by checking library websites manually and it was a time-consuming task.

Full-Text Indexing
Gale provided the Library with the metadata and full-text of articles. These were stored on the Trove servers. RMIT provided the metadata but not the full-text. Although it was initially hoped to full-text index all articles so that all words would be searchable, this was not possible with the existing infrastructure. To achieve the desired search performance with the available hardware, it was therefore decided not to full-text index everything.

Experimentation and testing was undertaken to try and find the optimal method for partial full-text indexing and to develop a stop word index. This resulted in a decision to full text index all articles that are less than 2500 characters. For articles longer than 2500 characters the first 2500 characters are full-text indexed (typically 500 words) plus up to 1000 unique non-stop words (or uncommon, or “characteristic words”) from the remainder of the article.
There are currently 633 words in the stop index. The testing seemed to indicate that abstracts and introductory paragraphs in the first 2500 characters usually describe the main content using the most common search terms people will use. Trove also fully indexes the article metadata so important concepts and search terms will typically be indexed as part of the assigned subjects, title and authors. This technique of partial full-text indexing stops the article index from growing too big and the heuristic is a trade off between good search performance and a perfect index. The identified deleterious effects for searchers of this type of indexing on subscription e-resources are that phrase searching only finds matches on the first ~500 words, and unique words from the ends of very long articles, particularly names of referenced papers and authors, will often not be indexed and hence will not be findable.

**Library holdings of Journals**
It had been hoped that Trove would be able to display only the library hard-copy holdings for specific issues or volumes of journals, rather than all the holdings for that journal. This would help users find the articles more easily. However, due to the inconsistency in metadata this was not possible, so the entire library holdings were linked to instead.

**Outstanding issues with linking and authentication to Gale articles**
- Some links to Gale articles fail. Persistent identifiers were not provided by Gale, so in most cases heuristics are used to map the Gale article ID into an effective search term. For the remaining cases, Trove links to the Gale OpenURL service which occasionally, due to differences between supplied metadata and metadata used to generate the Gale search indices, fails to resolve the article reference.
- Some Gale articles are hosted outside Gale and are accessed via links known as Digital Object Identifiers (DOI)s. Some of these links lead to full text, others only to citations. Some require authentication or payment and others are freely available. Trove is unable to identify which of these articles fall into which category.
- Gale occasionally has multiple copies of the same article in different products, and these may not have the same Gale ID or metadata. This means Trove does not merge the items onto the same work records.

**Outstanding issues with linking to Informit articles**
Informit allows subscribers who do not have their own authenticating proxy (such as EZProxy) to log in offsite using their library barcode. However, they are not automatically forwarded to articles. This makes the interface arrangement with Trove somewhat awkward, as Trove users click to log in to Informit, then must return to Trove and click again to view the correct article at Informit.

**Proxy Servers issues**
Two libraries had proxy servers which were unable to accept URLs from Trove. It was thought that this was something to do with the proxy server set up. These libraries were requested to liaise with other libraries using the same software and to contact their software vendors for advice.

**Vendor data updates**
Both Gale and RMIT were able to provide update files for each product containing new and deleted articles. Gale can provide these on a daily basis, whilst RMIT provides them at irregular intervals. Work was undertaken to enable Trove to automatically process the new additions in update files, but work to process deletions was slightly more complicated. There are estimated to be three million record deletions a year. Trove now automatically processes
deletions as they are received, along with new records in the update files.

**Vendor subscription updates**
Both vendors were able to provide updates to subscription data (e.g. new customers, changes to subscriptions, changes to IP addresses) on a regular basis. During the development stage the Trove team postponed the decision on how these updates would be handled. During development the details originally supplied by the vendor were manually added to the ALG, and it seems unlikely within the current infrastructure that updates could be automated. Work on this is planned for the future.

**Usage Statistics**
Clicks on links to full text hosted by RMIT and Gale are being logged within Trove. Work is planned for the future to make these statistics visible in Trove system reports. The number of searches on the new journals and articles zone is recorded within Trove Google Analytics reports, and will be able to be compared with other zone usage over time. The e-resource vendors will monitor referrals from Trove to measure increased usage of their articles and products via this new additional pathway.

### 7. FUTURE PLANS

The infrastructure for including subscription e-resource content in Trove has now been established. In the initial development stage two vendors and 120 million articles were included. The full extent of e-resources subscribed to by Australian libraries is unknown but is considerably larger. Identified work for the future is as follows:

**Trove system development work**
- Develop the ability to handle subscription updates from vendors, to add new customers, and apply cancelled subscriptions and changes in library details such as IP address.
- Develop a statistics report to monitor usage of subscription e-resources.
- Continue to improve the ‘get’ screens in Trove based on user feedback.
- Develop a Trove Application Programmers Interface (API). This would allow other libraries and organisations to send a search to Trove and bring back results sets which could be used to supplement the results in their own discovery service. The Library is now considering, in consultation with vendors, whether subscription e-resource content should also be offered via the API.
- Increase server performance and capacity for a 100% increase in content and usage over the next 18 months.

**Work with database vendors**
It is intended to expand the content in Trove by continuing to work with the existing vendors, as well as working with other vendors. Prioritisation of new content and vendors may be based on criteria such as: products that are heavily subscribed to by Australian libraries, are heavily used, contain Australian content, or are easy to process. Gale has offered another 50 products for inclusion and RMIT have offered the Informit Indigenous Collection.

The current inquiry into Australian school libraries, ‘School libraries and teacher librarians in the 21st century’, May 2011 recommends that the Commonwealth Government partner with all education authorities to fund the provision of a core set of online database resources,
to be made available to all Australian schools. If approved these e-resources may also be considered for inclusion in Trove.

**Work with discovery system vendors**
We may consider other approaches in the future, such as using APIs from discovery system vendors to obtain further content, or consider Google Scholar.

**Work with Australian libraries that have subscription e-resources**
In order for subscription e-resources to continue to work seamlessly in Trove, it is important that libraries know how authentication to the subscription e-resources works and how to keep their details updated.

If a library changes their proxy details they need to update their entry in the ALG. If a library does not have their proxy details in the ALG we assume that they are being authenticated by the vendor (many public libraries are set up this way with ‘barcode masking’). If a library changes its IP address this will be provided to the Trove team via the vendor subscription update files. Libraries need to supply the following information in the ALG for the authentication to work:

- Gale/RMIT customer id;
- Proxy server details. (If not supplied we assume the vendor is authenticating).

After updating the ALG the changes are reflected in Trove within half an hour. The Trove team explained to libraries what they need to do in a blog post written at the release of Trove 4.0. The link was widely circulated to libraries via listservs.

**Marketing Activities**
One of the biggest reasons cited by respondents of the NSLA survey in April 2009 for non use of subscription e-resources was that most users were simply unaware of their existence, or there were barriers to easily access them. By placing subscription resources in Trove there is a chance they may be more easily discovered.

Trove already has a high profile. However, although its usage is high and increasing, it is still largely focused on digitised Australian newspapers. The Library relies on capturing and educating users by viral marketing via user generated forums and blogs, twitter and occasional library blogs. There is still a need to make users aware of the existence of e-resources and their accessibility in Trove.

Trove key messages are being re-developed to reflect the new scope of Trove.

**8. MEASURING SUCCESS**

At the time of writing this article (September 2011) it was too early to evaluate the success of the development against the original goals:
- Give Australian library users improved access to e-resources, especially those e-resources subscribed to by NSLA member libraries and Australian public libraries.
- Ensure that subscription e-resources are used to the maximum extent possible, by supporting a user-centric approach to their discovery and access in the national discovery service Trove.
It is intended to measure the success by three methods in 2012:

- Feedback from Trove users, including feedback on any improvements in access to subscription e-resources.
- Measurements of trends in the number of e-resources accessed.
- A formal user satisfaction survey of Trove.

The possible success of the Trove development to include subscription e-resources largely depends on the following factors:

- Continued near-comprehensive contribution of data from Australian libraries, via Libraries Australia and other Trove contribution pathways;
- Continued support from RMIT and Gale Cengage in providing updates to subscription and data files included in Trove;
- Support from Australian libraries in notifying and updating, through the Australian Libraries Gateway, the details and changes to their authentication-related data;
- Support from Australian libraries in encouraging their users to use Trove, particularly for the discovery of e-resources and the discovery of collection holdings beyond their own Library;
- Continued work by the Library to maintain and expand Trove, including the addition of more e-resource content from the current vendor partners and from new partners.

9. CONCLUSION AND THANKS

It was a positive, rewarding and exhausting process to develop what had for so long been a vision, then a theoretical model, into a practical solution. Credit must be given to the commitment and passion of Dr Warwick Cathro who championed, led and encouraged the Trove team to reach its full potential and achieve these outcomes. Also to the partnership of Gale Cengage and RMIT without whom this would not have been possible. They were enthusiastic, helpful, and keen to work with us to achieve outcomes that benefit users. The input, support and collaboration of Australian libraries, in particular the NSLA Open Borders Project Group, greatly assisted the Trove team.

The National Library of Australia has made an ongoing commitment to maintain and support Trove into the future. It also intends to continue development and extend the inclusion of subscription e-resources in Trove.

The Trove team appreciated the ongoing support of all participating libraries and institutions for the development of Trove. It’s a service that would not be possible without the commitment of the whole of the library sector to collaboration and cooperation. We hope the inclusion of subscription e-resources builds on that to deliver benefits to all Australian libraries and their users. We have set a way forward for increased discovery, access and use of e-resources by Australians. Time will tell if we can call it a success.

10. REFERENCES


About the author:

**Rose Holley** has worked in libraries and archives in the UK, New Zealand and Australia. She is a digital library specialist and has managed a number of significant collaborative digitisation projects over the last 10 years, including the Australian Newspapers Digitisation Program. She now focuses on delivery of digital content to users, creating systems that enable social interaction and engagement with content. Rose is Manager of Trove the single discovery service for Australia based at the National Library of Australia. She has a particular interest in using crowdsourcing strategies to help improve access to library and archive collections, and in harnessing the knowledge, expertise and enthusiasm of volunteers.