Developing *Trove*:
the policy and technical challenges

Warwick Cathro
Assistant Director-General, Resource Sharing and Innovation
National Library of Australia
wcathro@nla.gov.au

Susan Collier
Project Manager, Trove Project
scollier@nla.gov.au

Abstract:
In September 2008 the National Library of Australia embarked on a project to develop a powerful new discovery service to expose the wealth of information in Australian collections. The new service, branded "Trove", was released in December 2009 after six months as a beta service. Trove is not only replacing eight legacy services, but is improving the discovery experience for the Australian public and researchers by including more content and by allowing users to engage with the content. This paper will describe the policy and technical challenges which were faced by the Library during this project, and will outline the Library's plans for the further development of Trove.

Background

One of the established activities of the National Library of Australia is its program of aggregating national metadata. This activity can be traced back to the establishment of the *Australian Bibliographic Network* in 1981, and in fact further than that – to the card-based union catalogues that were built in the 1960s. The underlying purpose of this program is to assist users to locate useful items in libraries other than their own, or (thinking beyond libraries) to facilitate the discovery and location of the wealth of information resources in the nation's collections.

Prior to the mid-1990s, these metadata aggregations were used by librarians as intermediaries for the user. The advent of the Web made it easier for these databases to be used directly by end users. In 1997, the Library released the *Register of Australian Archives and Manuscripts*, an online replacement for the printed *Guide to collections of manuscripts relating to Australia*. This was the Library's first "national online discovery service"; a free service aimed at assisting researchers and the general public to discover information resources held in Australian collections.

It was evident that a combination of collection digitisation and metadata aggregation could deliver even more powerful services for users. Accordingly, in 1999 the Library released *Picture Australia* as a service for discovering and accessing pictures that are digitised by Australian collecting institutions. After 10 years, that service provides access to over 1.7 million images from more than 100 separate collections. Those data contributors include museums, archives and other non-library institutions, an outcome facilitated by the use of a simple and generic metadata standard: Dublin Core.

Other national discovery services followed: *Music Australia* and *Australia Dancing* (2004), the search service for the PANDORA web archive (2005), and the *ARROW Discovery Service* (2005), now called *Australian Research Online*, which provides a single entry point for searching Australia's open research repositories.

In 2006, the Library realised a long-held goal by opening up the Australian National Bibliographic Database to the general public, through the *Libraries Australia* free search service. And in 2008, the Library launched its online service for searching and accessing digitised Australian historic newspapers. Thus, by late 2008 the Library was operating eight national discovery services.

During the period 2000 to 2006, most of these discovery services were built on a software platform called TeraText, supplied by the company SAIC, and formerly developed by RMIT University under the name SIM (National Library of Australia, 2002).

The integration imperative

By 2006 it was evident that this multiplicity of discovery services was not providing the best possible service for the user, and was hindering the Library's ability to maintain and improve the underlying software. In that year the Library undertook a review of its IT architecture, which reported in March 2007 (National Library of Australia, 2007). One recommendation of the review was that the Library adopt a "Single Business approach" to its digital library services, implying the development of a "single data corpus" which could be deployed in a range of contexts. The outcomes would be improved service for users (less aggregations to search) and more scope for the Library to maintain and innovate (by not having to manage many silo applications).

After a further planning process during 2007-08, the Library commenced this integration project in September 2008. The internal project name was "Single Business Discovery Project".

The aim of the project was to develop a discovery service that would:

- assist users by providing just one aggregation to search;
- reflect an emphasis on information resources held in Australian collections, without being entirely limited to such resources;
- provide access to a greater range of resources, including more full-text content which would be immediately available to users;
- enhance ease of discovery by providing features such as improved relevance ranking and search refinement;
- engage with users through content and metadata annotation services; and
- reduce the number of services requiring maintenance by the Library.

Late in the project this new service was given the name "Trove" – meaning a "treasure trove", defined in one dictionary as a "collection of valuable or delightful things". The name derives from the French "trouver", a verb meaning to find, or to discover. The name thus suggests the three concepts of a collection, of treasured or valuable collection items, and the process of discovery.

The Library decided to undertake this project as an in-house development, rather than use a vendor's product. In the Library's experience, vendor products provide satisfactory support for mainstream library functions, but the Library's preference is to use its own IT staff to develop more innovative digital library systems, usually based on appropriate open source software platforms. In this case, the in-house approach was further warranted by the unusual nature of the Trove data store, including digitised newspaper and web archive content.

Collection views

In the lead-up to the project, there was much discussion about whether users could be provided with a single relevance ranked result set covering all types of information resource. The project team concluded that it would be impracticable to deliver meaningful relevance ranking within such a single result set because of the different resource types and metadata schemas involved. Thus, Trove has been designed to deliver multiple result sets or “collection views”.

These “collection views” recognise the unique requirements of each type of information resource. Each view has its own home page, its own relevance ranking algorithm and its own facets, influenced by the type of material included in the view. Each view will define a set of “external targets” which will present additional results in those cases where important content or metadata could not be included in Trove’s own data store.

Many of these views are based on the format of the information resource (for example, “Pictures and Photos”). Others reflect a topic approach (for example “About People and Organisations”). The number of collection views, and the boundaries between them, may change as the scope of the service develops, and in response to feedback from users. In the future, there may be additional views reflecting disciplines (such as “Music”) which are represented in multiple resource formats, or views reflecting a wider contextual perspective (such as “Research”).

At an early point, the project team conducted a “card sorting” exercise with a focus group of potential end users. The exercise was aimed at giving the team a better understanding of the format categories that were understood by users, and of the terminology which would make most sense to users in describing those categories. The report of this exercise was published on the “Library Labs” site (National Library of Australia, 2008). The exercise was helpful in answering questions such as:

- do end users think of “books”, “pictures” and maps” as format categories?
- how do end users think of “academic papers?”
- do end users think of “archived websites” as a format category?
- do end users often group audio and video together?
- what do users understand by terms such as “archive” and “manuscript”?

The prototype

Following this exercise, the project team developed a prototype of Trove based on the following eight collection views:

- Books, Journals, Magazines, Articles…
- Pictures and Photos
- Australian Newspapers (1803-1954)

- Music, Sound and Video
- Maps
- Archived Websites (1996 – now)
- Diaries, Letters, Archives ...
- About People and Organisations.

The prototype was released in May 2009. It included a prominent “feedback” box, and in the following six months more than 600 comments and suggestions were received. Many of these led to improvements, primarily to the interface design, and bug fixes.

The prototype attracted some international recognition. One prominent commentator described it as “one of the best one-stop shopping discovery portals I've seen” (Tennant, 2009).

Content and coverage

The wide coverage of Trove has been assisted by the Library’s activities in areas such as:

- managing Libraries Australia;
- building the PANDORA web archive;
- previous involvement in the ARROW Project; and
- undertaking the digitisation of Australian newspapers.

The coverage can be illustrated by the following concrete example. Suppose that a scholar is researching the life and works of Ethel Turner, the author of “Seven little Australians”. Through a single search of Trove that scholar would be able to access the following information which would be presented on a single web page, grouped into several result sets:

- books by and about Ethel Turner, with information on the location of those books in Australian libraries, and with access to the full content where the work is out of copyright;
- articles, conference papers, theses and other research dealing with Ethel Turner, including content from university open access repositories and articles in e-journals (and in the latter case, with support for navigation to the full content where the scholar’s library has a subscription to a product containing the e-journal);
- pictures of Ethel Turner from libraries, museums and archives, including digitised pictures and information about the location of pictures not yet digitised;
- newspaper articles dealing with Ethel Turner, and published prior to 1955;
- archived web sites that refer to Ethel Turner;
- music, sound and video resources, including audio books and information about the ABC television series of *Seven little Australians*;

- information about papers, letters, diaries and other records relating to Ethel Turner that are in archival collections; and
- biographies of Ethel Turner from sources such as the Australian Women's Register, the Dictionary of Australian Biography Online, and Wikipedia.

Structured metadata sources used by Trove include:

- Bibliographic and authority records from the Australian National Bibliographic Database (ANBD);
- Dublin Core metadata from Australian Research Online (ARO) and Picture Australia (existing National Library discovery services);
- Dublin Core metadata from OAIster, a union catalogue for worldwide university and other academic repositories, maintained by the University of Michigan until October 2009, and thereafter by OCLC;
- Metadata from the Open Library (http://openlibrary.org/), an Internet Archive project which aims to create a web page for every book in the world – Trove harvests only those records that link to full text;
- Metadata from the Hathi Trust (http://www.hathitrust.org/), a digital archive of library materials converted from print that is co-owned and managed by a number of academic institutions – Trove harvests records only for works that are in the public domain; and
- Tags from Wikipedia (where an ISBN is found in Wikipedia, the title of the Wikipedia article is used as a tag).

Trove also includes full text in which every word is indexed. The key sources of full text include:

- Manuscript finding aids – currently only those from the National Library;
- a selection of 120,000 books from the Internet Archive;
- about 1400 e-books from the University of Adelaide (http://ebooks.adelaide.edu.au/) (a collection of classic works of literature, philosophy, science, and history); and
- descriptions, sample chapters and tables of contents for 400,000 books, harvested from the Library of Congress.

Biographical data

The inclusion of biographical data in Trove reflects the fact that the Library has been building a data contribution program called “People Australia” (Dewhurst, 2008). This program aims to construct a virtual web page for each person or organisation in this data aggregation, and to support navigation to biographical information contained in other web-based services.

In parallel with the development of Trove, the project team developed software to:

- harvest the biographical data and to convert it into a schema compatible with the Encoded Archival Context (EAC) schema (http://eac.staatsbibliothek-berlin.de/);
- support the matching of harvested names with those already in the database, and support action by reviewers to resolve uncertain matches; and
- expose the biographical data through standard machine-to-machine protocols (OAI, SRU, OpenSearch).

By September 2009, the following biographical data had been harvested:

- the Australian Name Authority File;
- the biographical data in Music Australia and Australia Dancing; and
- the Australian Women's Register.

Many other potential contributors have been identified, including the Australian Dictionary of Biography Online, the Dictionary of Australian Artists Online, and Bright Sparcs (which documents Australian scientists). At the time of writing, the National Library was in discussion with the Australian National Data Service (ANDS) concerning a possible project to extend the People Australia program by harvesting information about Australian university researchers.

This program is raising some data issues, given the sparse information in the Name Authority File and the existence of some duplicates in that file. In addition, many important Australians are excluded from the data harvested so far. For this reason, Trove will include extracts from biographical entries in Wikipedia, with links to the full Wikipedia entries. Wikipedia is considered to be a source of reasonable data quality and, although less authoritative than (say) the Australian Dictionary of Biography Online, it is a comprehensive and up-to-date source suitable for supplementing the harvested biographies.

Technical issues and challenges

Software

Trove was developed in-house by the Library, using the Java programming language. Trove uses Lucene for indexing and searching both metadata and full text, and Solr, a web service that makes commonly used features of Lucene easily available. The project chose these tools because they are open source, and because Lucene is fast, flexible, reliable and scales very well. The Library had already used these tools successfully in other projects (the development of Australian Newspapers, and the renovation of Australian Research Online). These tools have a large and active developer community.

Trove also uses:

- MySQL as the central metadata store (http://www.mysql.com/);
- Restlets as the service framework (http://www.restlet.org/);
- Jetty as the HTTP container (http://jetty.codehaus.org/jetty/); and
- FreeMarker as the templating language (http://freemarker.org/).

Architecture and data harvesting

The overall system architecture and data flow for Trove are summarised in Appendices 1 and 2. Note that Trove uses four Lucene indexes: a main index, and separate indexes for the PANDORA web archive, the newspaper content, and the People Australia content.

Currently Trove is relying on two OAI-compliant data harvesters. One of these, the Picture Australia harvester, is based on the MPS software from SAIC (http://www.teratext.com/products/teratext-metadata-publishing-system.asp). The other ("NLA Harvester") was written by the Library, and is being used to harvest the data for the Australian Research Online and People Australia programs. Once the NLA Harvester has been extended, it will replace the Picture Australia harvester.

Currently Trove does not have enough capacity to index every item of full text which it could harvest from the sources described above. At the time of writing, the equivalent of about 120,000 full-text books had been indexed.

Hardware configuration

To provide redundancy, Trove uses two production servers designed to mirror each other. Normally they share the load. If one server fails, Trove is able to continue running on the other server.

To improve search performance, Trove uses four Solid State Drives (SSDs) in each server. When the size of the active part of the Lucene index greatly exceeds the available memory size (search is accessing an index of approx 400 GB) performance is degraded. Instead of buying more memory, the Library has chosen to use SSDs, as very fast and high quality SSDs are now very cheap relative to memory.

Update performance

One downside of Solr is that it is relatively poor at supporting systems where record updates need to be quickly visible. This has been a challenge, because it is desirable for Trove to incorporate changes such as user tags and comments in near real time. As at January 2010, it can take up to one minute between making a change such as adding a tag and having it visible in the interface. The project has judged this delay to be acceptable at present, and by mid 2010 hopes to benefit from the real-time update work being done as part of the Lucene project (Apache Lucene, 2009).

Grouping of Works and Versions

It will be recalled that one aim of the project was to make discovery “as easy as possible” for the user. Part of this involves making the presentation of result sets as intelligible as possible for the user. A framework supporting this aim is FRBR (Functional Requirements for Bibliographic Records) which uses the hierarchy Work, Expression, Manifestation, Item (IFLA, 2009).

The Trove result set display does not implement the full FRBR model, but focuses on grouping records into Work and Version clusters. A Version can be considered as a level intermediate between Expression and Manifestation.

Trove attempts to group each record from any of its contributing sources with other records in the database. When a user searches the system, each grouped record appears as a single search result. The rules used to group records into a Work cluster are as follows:

- A record is classified into a format such as book, picture, or map. Only records of the same format type can be grouped together.
- A scoring system is used to determine whether two records are a match. Records are assigned a score for each match point found. If a threshold score is reached, the records are grouped together. Match points include:
  - OCLC Work Identifier (this is considered a definitive match)
  - Authors and Creators
  - Title, Uniform title, Other titles
  - ISBN, ISSN, LC Control Number
  - Publisher (very low scoring).

For example, "Alice in Wonderland" the talking book, by Lewis Carroll, would be grouped in the same Work cluster as an ordinary text version of the same book, but not with a video of "Alice in Wonderland". The edition of "Alice in Wonderland" by Lewis Carroll, published in London in 1966 would be grouped in the same Work cluster as the edition published in America in 1982.

A Work cluster is further sub-sorted into Version clusters. Language, sub-format, publication dates, edition statement and/or numbers, publication details, physical description, author and series are used to assign a record to a Version cluster. If all these fields match in more than one record, then the records are grouped into a single Version cluster.

Each Version cluster will potentially record multiple “Items” held by multiple libraries.

Since these matching algorithms can never be perfect, users have been given the ability to amend the Work and Version clusters. A user can merge one work (Work A) into another (Work B), in which case Work A becomes a Version within Work B. Conversely, a Version can be separated from a work to become its own work. The user is able to add a note to each change to explain their reasoning for the move. To date a small number of interested users have used this feature.

Online access

It is important that Trove allows users to limit their searches to “online resources” but it has proved difficult to determine what level of online access is provided from a URL in the metadata record. In some cases this URL links only to a table of contents or a publisher’s description, and in other cases a subscription is needed to enable access to the full resource. The project team undertook extensive analysis to identify clues to the level of access. Trove supports three categories of link: “available online”, “available on line (access conditions apply)”, and “possibly available online”.

User interaction

Australian Newspapers, which was released as a beta service in July 2008, provides for significant interaction by the user community, including the ability to correct the OCR text, and to add tags and comments. Trove will extend this kind of user interaction to all collection views.

Users will be able to add tags or comments to any record, and this user-contributed data will be searchable in Trove. In addition, as we have seen, users will be enabled to move a record from one collection view to another, and to amend the work and edition groupings.

Future developments

The National Library has set out its plans for the future of Trove in the Strategic Plan for the Resource Sharing and Innovation Division (National Library of Australia, 2009).

The Library implemented the first stage of Trove during December 2009. This implementation was accompanied by the decommissioning of two legacy software applications as stand-alone services: the Libraries Australia free search service, and the Register of Australian Archives and Manuscripts. Another stand-alone service, the Australian Newspapers service, will be decommissioned in the first quarter of 2010.

The Library is now working on further improvements to Trove. As at January 2010, it is making a number of general improvements such as provision for users to create lists for a range of purposes, RSS feeds, enhanced sorting of results, more external targets, indexing of more full text, further improvements to the user interface, and bug fixes. The Library then plans to develop the Australian Newspapers collection view to meet priority enhancements identified during the Australian Newspapers Beta phase.

During 2010, the Library plans to commence the decommissioning of other legacy services, commencing with Picture Australia. This decommissioning will be actioned only when Trove has at least matched the user interface and functional features of the legacy service, including the provision of trails and advanced searching.

In the second half of 2010, the Library plans to expand the journal article content in Trove. This will involve:

- support for search and delivery of the full content of any journals digitised by the Library;
- inclusion of journal article indexing data and the linking of this data to library holdings; and
- the ability for Australian library users to easily discover and link to e-resources, from selected vendors, which they are entitled to access by virtue of their library memberships.

The Library aims to strengthen the coverage of Trove by exposing content held by archives and museums. The preferred strategy is to obtain high level collection guides, finding aids and file descriptions, to index the text of these guides, and to rank them highly in the “Diaries, Letters, Archives” search results. The user will navigate from the copy of the guide in Trove to the original guide at the archive or museum web site and, where possible, to specific items within the collection, where these are documented on that web site. This strategy will effectively provide a context for the user and an appropriate way of linking to archive and museum collection content.

The Library plans to work with State and Territory libraries to identify additional content for Trove, and to identify any opportunities for these libraries to build new discovery services of their own by leveraging off the Trove data collection and its application programming interfaces.

Conclusions

The release of Trove follows three years of discussion, analysis and development by the National Library, a process which began with an IT Architecture review in 2006. Trove has taken the Library’s discovery services to a new level, in terms of the degree of integration, the expansion of full text content, and the extension of user contribution to content categories beyond newspapers.

Many integrated portals have been developed by libraries, library networks and vendors. Trove is unusual in its breadth and in its national scale, given its inclusion of archived web sites, digitised newspaper articles, university research outputs, the national union catalogue and online biographical databases.

References

Apache Lucene – overview (2009). http://lucene.apache.org/java/docs/index.html

Dewhurst, Basil (2008). People Australia: a topic-based approach to resource discovery. http://www.valaconf.org.au/vala2008/papers2008/116_Dewhurst_Final.pdf

IFLA (2009). Functional requirements for bibliographic records.
http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records

National Library of Australia (2002). Media Release. 12 August 2002.
http://www.nla.gov.au/pressrel/2002/teratext.html

National Library of Australia (2007). IT architecture project report
http://www.nla.gov.au/dsp/documents/itag.pdf

National Library of Australia (2008). Single Business card sort report, December 2008. Linked from https://wiki.nla.gov.au/display/LABS/1.+Our+Prototypes

National Library of Australia (2009). Resource Sharing and Innovation Division. Strategic Plan, July 2009 to June 2011.
http://www.nla.gov.au/librariesaustralia/documents/div3-strategic-plan.pdf

Tennant, Roy. One-stop searching with a can-do attitude. 26 May 2009.
http://www.libraryjournal.com/blog/1090000309/post/1900044990.html

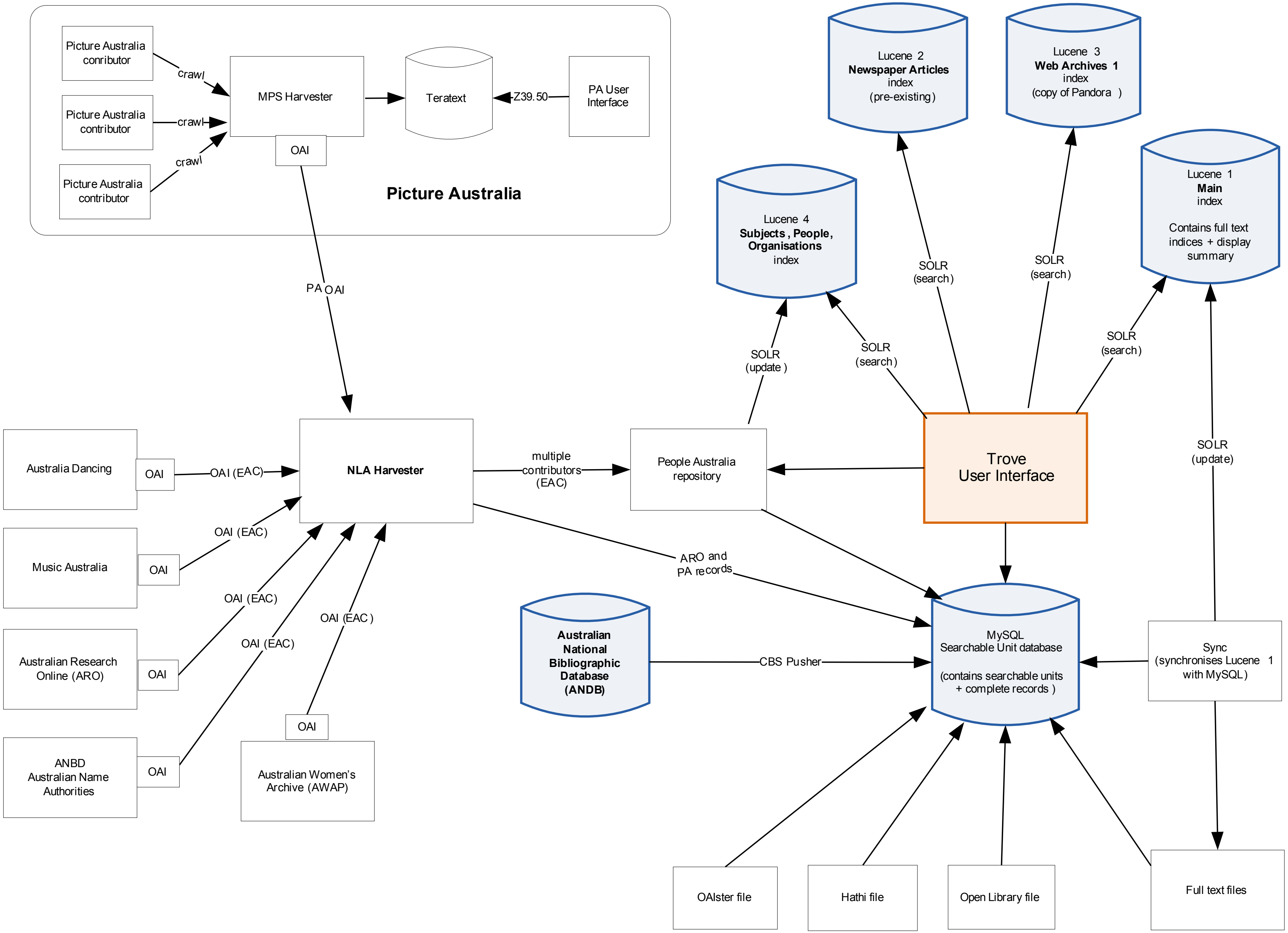
Acknowledgments

The authors wish to acknowledge the support and advice of many current and former National Library staff during the development of Trove and in the planning phase prior to the project.

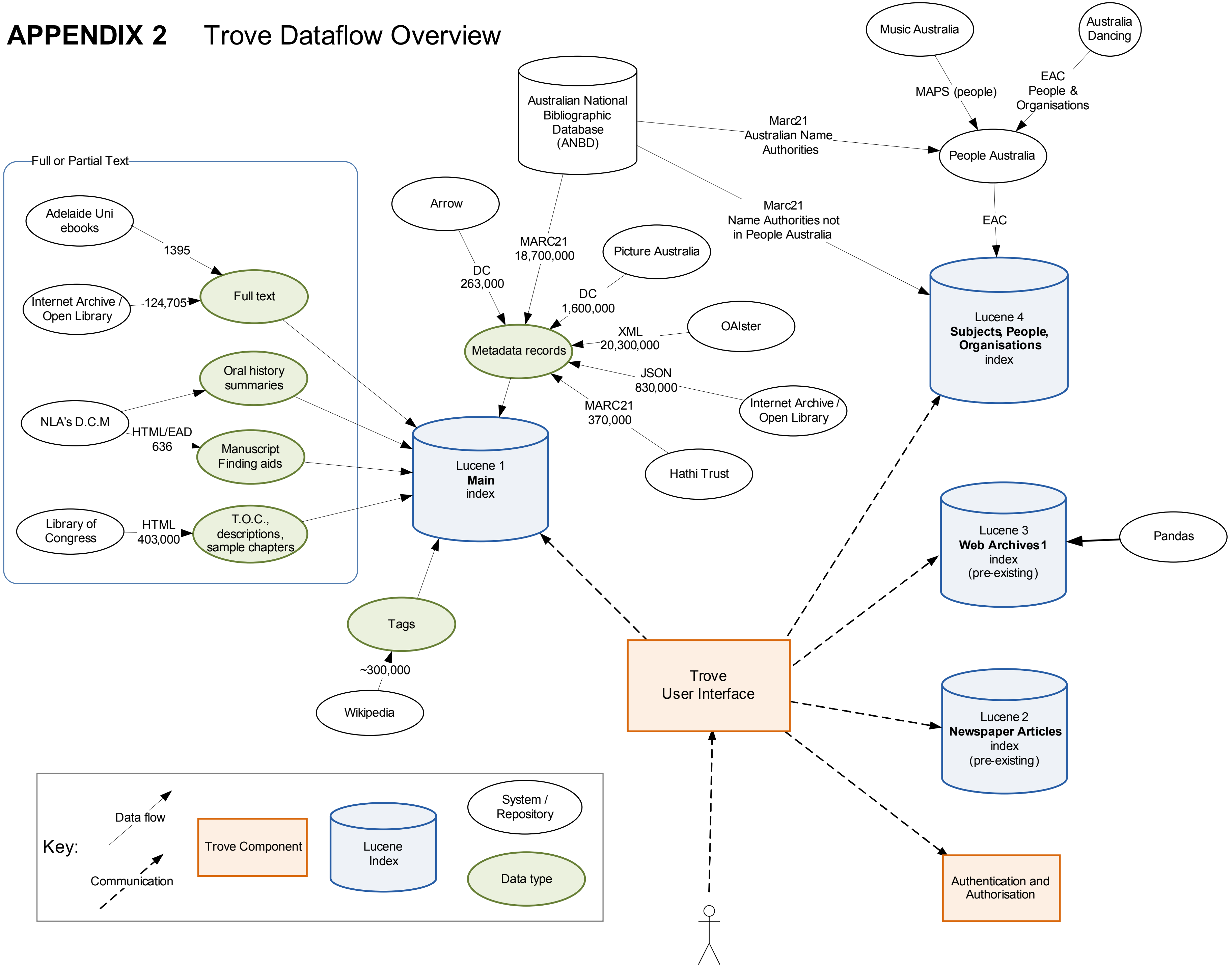
In particular, the authors wish to acknowledge the contributions of the other members of the Trove Project Team: Kent Fitch, Joanna Meakins, Simon Jacob and Kenji Walter.

The authors also thank the many users and members of the public who provided comments on the prototype of Trove.

APPENDIX 1 Trove Architecture Overview



APPENDIX 2 Trove Dataflow Overview


Music Australia
Australia Dancing
MAPS (people)
EAC People & Organisations
Australian National Bibliographic Database (ANBD)
Marc21 Australian Name Authorities
People Australia
Marc21 Name Authorities not in People Australia
EAC
Full or Partial Text
Adelaide Uni ebooks
1395
Internet Archive / Open Library
124,705
Full text
Oral history summaries
NLA's D.C.M
HTML/EAD 636
Manuscript Finding aids
Library of Congress
HTML 403,000
T.O.C., descriptions, sample chapters
Arrow
DC 263,000
MARC21 18,700,000
Picture Australia
DC 1,600,000
Metadata records
XML 20,300,000
OAIster
JSON 830,000
Internet Archive / Open Library
MARC21 370,000
Hathi Trust
Lucene 1 Main index
Lucene 4 Subjects, People, Organisations index
Lucene 3 Web Archives1 index (pre-existing)
Pandas
Lucene 2 Newspaper Articles index (pre-existing)
Tags
~300,000
Wikipedia
Trove User Interface
Authentication and Authorisation
Key:
Data flow
Communication
Trove Component
Lucene Index
System / Repository
Data type