Access to online archival catalogues via web APIs

Richard Lehane, State Records NSW

This paper describes State Records NSW's Open Data Project and the development of a web API to the online catalogue, Archives Investigator*. In the course of developing the web API, a new experimental user interface to the catalogue was also built. Key features of that new interface are described.*

State Records NSW Open Data Project

Archival descriptive data has its primary role in producing online catalogues or finding aids. But it has significant secondary value, too. Some of this secondary value is already being realized: for example by permitting Google and other search engines to index our catalogues we enable the secondary use of descriptive data in search results. Many archives engage in a more deliberate sharing of descriptive data by collaborating with other cultural institutions to create federated search portals.

Riley and Shepherd (p 94) make the case for an even wider sharing of archival descriptive data: 'In many cases, shared archival descriptive metadata can be of use to those with whom archives have no pre-existing relationships, to create "mashups" combining archival metadata with that from other (including commercial) sources.' Riley and Shepherd imagine cultural hackers joining archival descriptive data with services like Google Earth or Wikipedia to create new perspectives on archives.

This idea, that data created for one purpose can potentially be re-used in all sorts of creative ways, of course goes much wider than just archival data. In 2009, Tim Berners-Lee called on scientists, governments, and institutions to publish their 'raw data' online. This call has been taken up by governments, sparking an 'Open Government Movement' and the creation of online clearing-houses of government data such as http://data.gov and http://data.gov.uk. In New South Wales we have http://data.nsw.gov.au. These initiatives seek to improve government transparency and enable the innovative re-use of government data by developers. An example of such citizen-led innovation is FlyOnTime (http://flyontime.us), a website that combines data from US federal

government agencies with weather and user-generated data to give an accurate picture of flight waiting times and travel conditions.

Inspired by the Open Government Movement and the potential re-use value of archival descriptive data, State Records NSW initiated its own Open Data project (http://data.records.nsw.gov.au) in March 2011. The aim of this project is to identify datasets relating to the NSW State Archives collection and publish them in accessible ways. It is envisaged that data published by this project could spark new interfaces to the collection, create new possibilities for federated searching, or allow creative re-purposing such as in visualisations or mashups.

As a first step, the project published raw data extracted from State Records NSW's online catalogue, *Archives Investigator*. This was made available as XML files representing each of the descriptive entities in the catalogue (State Records NSW's archival control system is structured according to the Australian Series System and its archival descriptions are formed from relationships between entities: record items, record series, agencies, persons, organisations, ministries, functions and activities). The data was also provided as a single SQLITE database file to represent the whole catalogue. These data files were accompanied by a web blog describing the data.

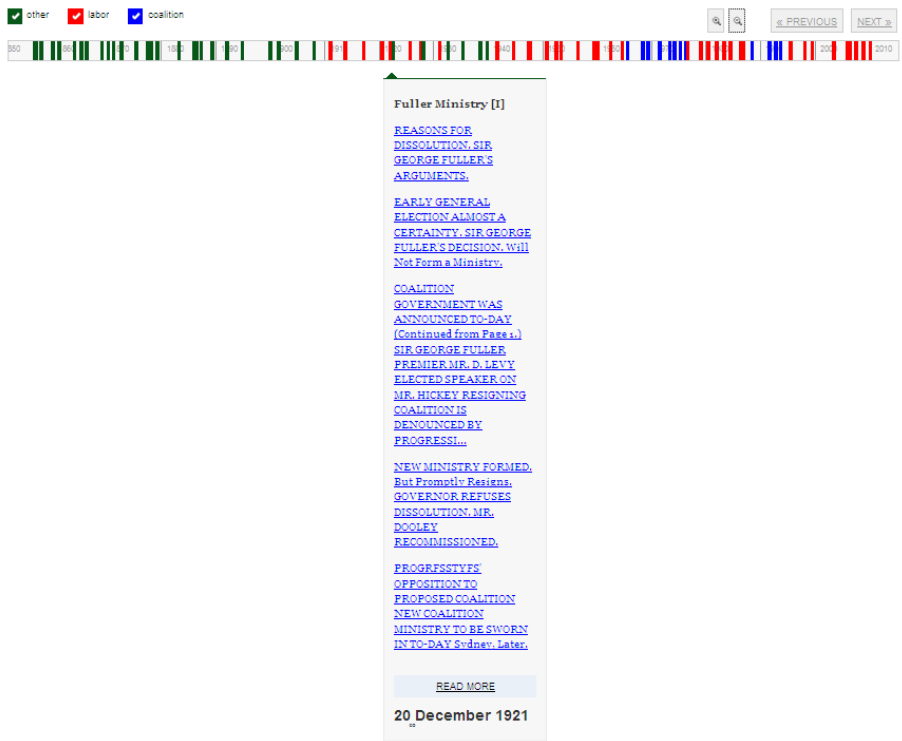

Fig. 1 Sample mashup – timeline of NSW government ministries

Publishing data is one thing, seeing that data used is another thing entirely. In order to promote re-use by showcasing possibilities, a sample mashup was created combining the data for government ministries with a visual timeline, Wikipedia, and Tim Sherratt's experimental API to the National Library of Australia's Trove Newspaper service. This example has proved a useful tool in explaining the purpose of the project. However, a failing of the project is that very little independent re-use of the data has so far occurred.

Building a web API

The project's next step was to build a Web Application Programming Interface (web API) to the *Archives Investigator* data. Web APIs are simply interfaces that developers can use to write applications that connect with your service. If you've used a Twitter client or a custom Flickr application, you've indirectly benefited from the web APIs that both of those companies provide.

Broadly speaking, the API provides access to the 'nouns' of State Record's catalogue (the entities: series, items, activities, agencies, etc.) and to the 'verb', *search*. Entities can be accessed either singularly or as lists through logical URLs (e.g. http://api.records.nsw.gov.au/agencies/1.xml or http://api.records.nsw.gov.au/agencies.xml) and are available in multiple formats (XML, JSON and some additional standard formats such as EAC CPF). Search results are available in XML (the OpenSearch format) and JSON. Key protocols like OAI-PMH have also been implemented.

Amanda Hill (p 146) raised the benefits of web APIs for online archival catalogues over eight years ago (referring to 'web services'):

> By allowing other computer systems to interact directly with our finding aids we are opening up possibilities of presenting archival data within any number of other applications and portals. These could be a world-wide archival network, a subject specific gateway, a corporate or institutional portal, or a local search service.

A web API to the catalogue obtains the same benefits as making the raw data available for download. It also has a number of distinct advantages:

- it is more timely, the data updates whenever changes to the catalogue are made
- data can be made available in a variety of formats through a web API (such as Dublin Core, EAC CPF, and MODS)
- a web API can provide access to the catalogue's functionality (searching, tagging, commenting, etc.) as well as to its contents
- external users no longer have to download the full dataset to make use of it, they can selectively access the particular descriptive entities they are interested in.

As well as benefitting State Records NSW's external stakeholders, the web API provides a flexible and open platform that State Records NSW itself can leverage to create new and innovative online services, such as mobile applications. For example, State Records NSW recently trialled a Flickr group where users were encouraged to upload their own images of archives to share with others. The web API was integrated with this project so that when an uploaded image was tagged with a series or item identifier it was automatically displayed alongside that catalogue description.

It is hoped that the web API can continue to be developed and, in the future, expand to encompass more of State Records NSW's core business functions. Barbara Reed has sketched out a vision of 'service oriented architectures and recordkeeping' in which organisations implement recordkeeping processes through standalone, interoperable services. This would allow, for example, an archives authority to publish a disposal scheduling service to which government agencies subscribe. There is certainly scope for State Records NSW's web API to extend to services, like disposal authorisation, classification, and transfer and access to digital archives, where responbilities are shared with agencies and where there are opportunities for greater automation.

‘An experimental new search tool’

Building the web API involved developing a whole new web frontend and search index for *Archives Investigator*. Because of this, it was natural to also create a new user interface for the NSW State Archives Collection, it was just a matter of adding new HTML ‘views’ of the data alongside the XML and JSON representations. This new user interface has been promoted to State Records NSW’s user base as an 'Experimental new

search tool'. Oddly enough, this offshoot of the project has proved to be one of its greatest successes and it has meant that everyday users have benefitted and become champions for the project.

For regular users, the new user interface offers simple but powerful search, a richer search results page, a cleaner view of descriptive entities and their interrelationships, and integration with external tools such as Zotero (http://www.zotero.org). It has also enabled experimentation with new features such as user tagging and commenting.

Improving search

A core challenge in providing search access to archival collections is that users will overwhelmingly choose to use simple search options when advanced search strategies typically yield much better results (because items are often under-described so won't appear in results and because subject-type queries can often only be answered by identifying relevant related agents or functions). In many cases users depend on archivists in reading rooms to reframe their queries in archival terms: i.e. by asking what creator or business activity might have generated records relating to the user's query. Of course for users unable to attend a reading room the problem remains.

In the 'Experimental new search tool' we address this core challenge by offering only simple search and generating an 'advanced' results page.

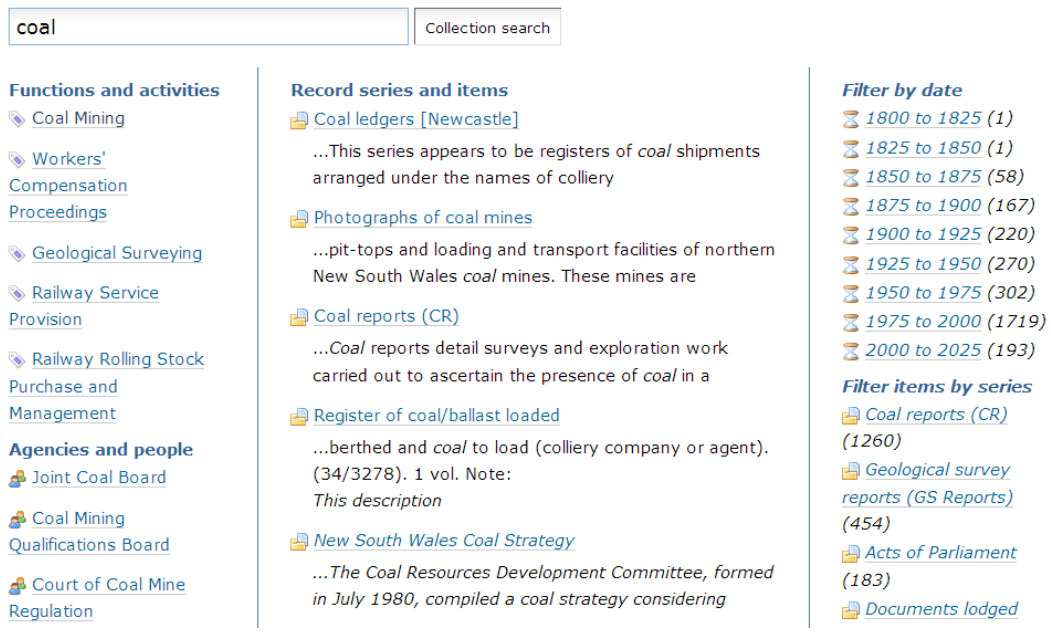

Fig. 2 Search results page

What makes it an advanced results page?

One important feature is the search filters (on the right side of Figure 2). These allow users to narrow their results based on date range, record series and location. These kind of choices are typically offered in an advanced search form but they work better as filters on results because, rather than crafting a very specific query at the outset that may return no results, users can ask a broad query returning a large set of results and then iterate on it, 'zooming in' to find the best material.

The other key feature of the results page is the structured view it provides of descriptive entities. Rather than presenting search results as a simple list, the new search provides a structured view, clustering results according to three questions:

- **what** records (both record series and individual items) relate to the query?
- **why** might records relating to the query have been created by government (government functions and activities)?
- **who** in government (agencies and people) might have created records relating to the query?

In other words, the search engine itself reframes the user's question into an archival one. The three-part division of descriptive entities is consistent with Chris Hurley's conception of archival description as comprising three essential types of entity: documents, deeds and doers.

Why is this structured view of the results better than simply providing links to series or items? After all, we know that, if asked, users will say that they just want a search engine to take them straight to the 'right' item? Well, for one thing, many queries (especially on general topics such as 'war' or 'unemployment') will not return good results if limited to only series or items. In such cases, exploring contextual pathways can yield much better results. In any case the contextual entities aren't forced on users: series and items get the most prominent treatment on the page and users can ignore the **why** and the **who** if they wish. By presenting those entities, even to an initially uninterested audience, it is hoped that users will intuitively attain a better understanding of the archival descriptive model. David Bearman (p. 45) urges archivists

to construct, 'a model of the archives as an information system, which users can maintain as an archetype and employ to navigate through the documentation which archivists create.' A structured view of results serves as such an archetype. It might also help users understand the records better too. When thinking about how to present archival context online we have an unfortunate tendency to focus narrowly on just the display of contextual information on the ultimate page on which a record (or information about a record) is presented, a kind of a contextual 'heads-up display' that frames the record with meaning. We must also remember that in the journey to that ultimate page, during the process of discovery and of navigation, context accumulates in the minds of users.

Better views of descriptive entities

In developing the new user interface, attention was also given to the way the catalogue's descriptive entities are presented.

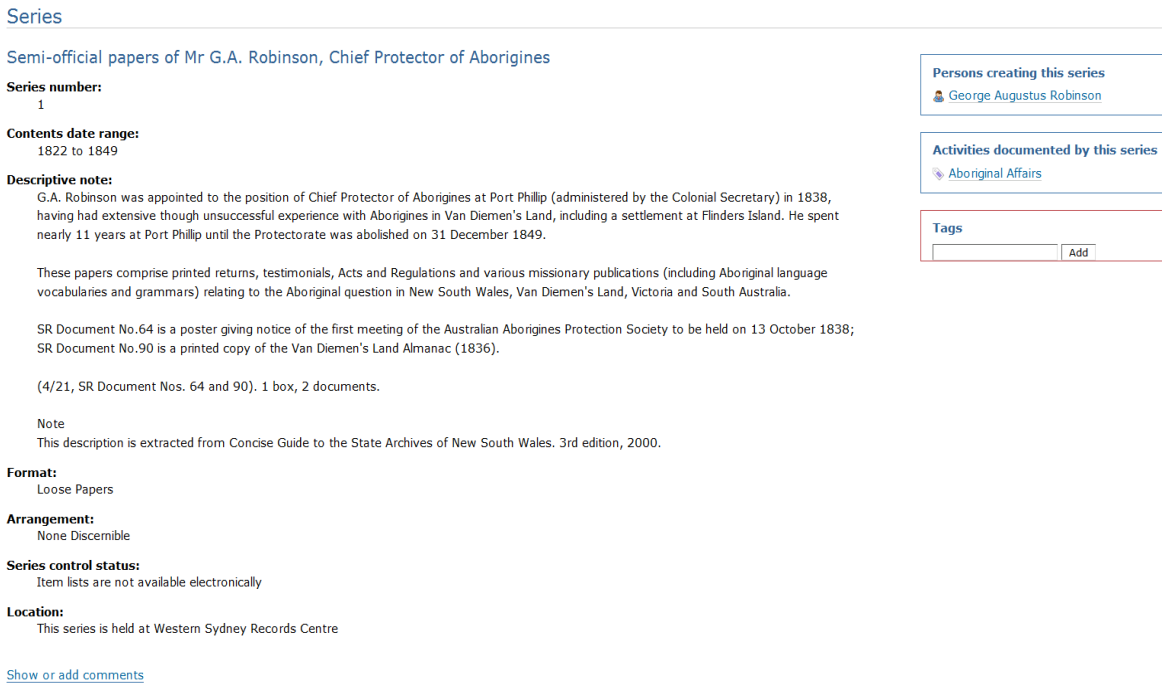

Fig 3. Example series description

The two objectives in this re-design were:

- to simplify the presentation of information by removing non-essential fields and by consolidating fields wherever possible (e.g. separate start and end date fields

become a single date range). An entity description should not look like a database report.

- to give greater visual prominence to relationships with other entities.

Relationships are at the heart of the Australian Series System. In State Records NSW's legacy catalogue, however, they were placed at the bottom of the page and were often practically invisible, hidden 'below the fold'. By positioning them in a dedicated column, it is hoped that users will spend much more time exploring these links. Wikipedia is a rare site that sucks its visitors in: you arrive looking for one thing and leave much later having browsed between entries, following a trail of connections. A Google-like ease of searching is great but, if users are to make sense of collections and unlock the richness of archival descriptions, then we should seek to emulate Wikipedia's stickiness too.

Tagging and commenting

Tagging and commenting features were also added to the new user interface. However these features have not been very successful, finding little use.

Tags have had the indirect benefit of allowing user generated content from State Records NSW's Flickr groups to be pulled in to the catalogue. They have also proved useful for developers using the web API. For example, the *Invisible Australians* project has tagged records relating to Chinese Australians for automatic extraction (http://invisibleaustralians.org). Tags and comments can be added automatically through the web API (it is read-write). Hardly any comments have been written to the site.

This failure to foster significant engagement through tags and comments can be attributed to a lack of resourcing for community building (promoting these features and moderating the content). You can't just tack Web 2.0 features on if you want them to work, build it and they won't necessarily come!

Conclusion

The Open Data Project and State Records NSW's web API continue to yield benefits for the organisation. The web API has proved very useful as an open infrastructure for sharing descriptive data. For example, State Records NSW was recently asked to provide series descriptions in a custom XML format for pooling in a research data catalogue. A request of this type would have previously required a dedicated (and resourced) ICT

project. With the web API, we could provide an initial test export within hours. The Open Data Project has also provided a wonderful opportunity for experimenting with the shape of our catalogue. Online discovery of archives is not a solved problem and we can't just rely on generic search solutions to address it. Not all of State Records NSW's experiments have proved fruitful, but we have made improvements, especially in the structured presentation of search results and in the site's 'stickiness'.

References

David Bearman, 'Documenting Documentation', *Archivaria*, Vol. 34, 1992

Daniel Bennett and Adam Harvey, *Publishing Open Government Data*, W3C Working Draft, 8 September 2009, available at http://www.w3.org/TR/2009/WD-gov-data-20090908/

Tim Berners-Lee, *Tim Berners-Lee on the next web*, TED, 2009, available at http://www.ted.com/talks/tim_berners_lee_on_the_next_web.html

Amanda Hill, 'Serving the Invisible Researcher', *Journal of the Society of Archivists*, Vol. 25, No. 2, 2004, pp 139-148

Chris Hurley, *Documenting Archives and Other Records - A Guide for Dummies*, August 2008, available at http://www.infotech.monash.edu.au/research/groups/rcrg/publications/ch-documenting-archives-supplement.pdf

Barbara Reed, 'Service Oriented Architectures and Recordkeeping', *Records Management Journal*, Vol 8, No. 1, 2008, pp 7-20

Jenn Riley and Kelcy Shepherd, 'A Brave New World: Archivists and Shareable Descriptive Metadata', *The American Archivist*, Vol. 72, 2009, pp 91-112