

Conal Tuohy
ConalTuohy.com
@conaltuohy
conal.tuohy@gmail.com

Summary

Metadata is of interest in its own right

Metadata are hugely informative, and can be research datasets in their own right, independently of any value they may have in supporting resource discovery. In general, data collections have to be seen as multivalent; possible readings and uses are truly open-ended.

"Distant Reading" is a new and important scholarly method

We should expect a continuation of the growth in "Distant Reading" (the term coined by Franco Moretti).

To support distant reading, we need to provide scholars with automated access methods to collections *in bulk*, not just to individual items mediated by some discovery interface. Wholesale, not just retail.

- Image collections: these can be "distantly viewed" using automated image analysis, face recognition, etc. software.
- Text collections: can be distantly read using topic modelling, named entity recognition, etc. tools.
- Metadata collections: can be distantly viewed with statistical and network analysis methods.

Capturing value of scholarly research for the library

New uses of library collections can generate value for the library, e.g. generating new metadata for discovery. How to mobilise and capture that new metadata for the library?

Libraries which provide tools for scholars to analyse and annotate their collections are in a position to capture that added value.

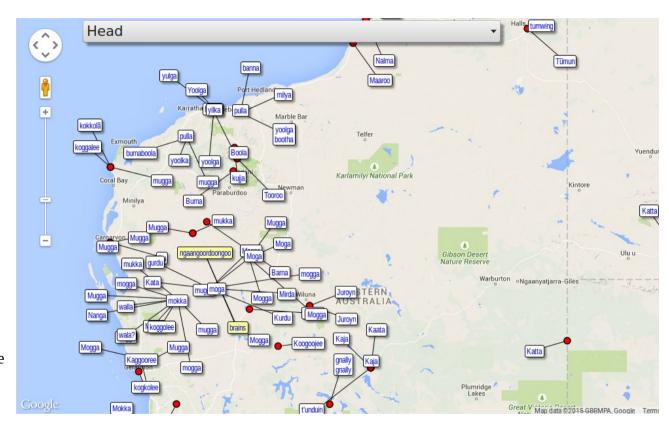
Example 1. Daisy Bates

http://languages-linguistics.unimelb.edu.au/research/projects-daisy

This example shows how a textual archive held by the National Library can be treated as a complex dataset, serving as the basis for computation and geographic visualisation.

There are 4,500 pages of typescript and 8,600 pages of manuscript, representing languages from the Southern SA/WA border up to the Kimberley. At least 123 speakers are named in the vocabularies and it is unclear how many languages they represent.

The 4,500 pages of typescript and 1712 notebooks in the NLA will be imaged, keyboarded, and the text will be encoded using the TEI framework. We will base our approach on work done by Henderson (2008) on part of the Gerhardt Laves manuscripts at AIATSIS. Bates' work represents a resource that has not been made accessible, and not even the most basic



work has been done to ascertain what languages are represented (eg a best guess approach was taken in Thieberger 1993; Nash 2002 applies a series of metrics to sample data from south-east WA to show similarities to known languages of the region, and McGregor 1998 has been able to identify some Kimberley languages). At a time when Australian languages are under severe threat from English it is critical to make the best representation of historical sources to assist speakers in their efforts to relearn and reinforce the use of their languages.

Example 2: Newspaper research

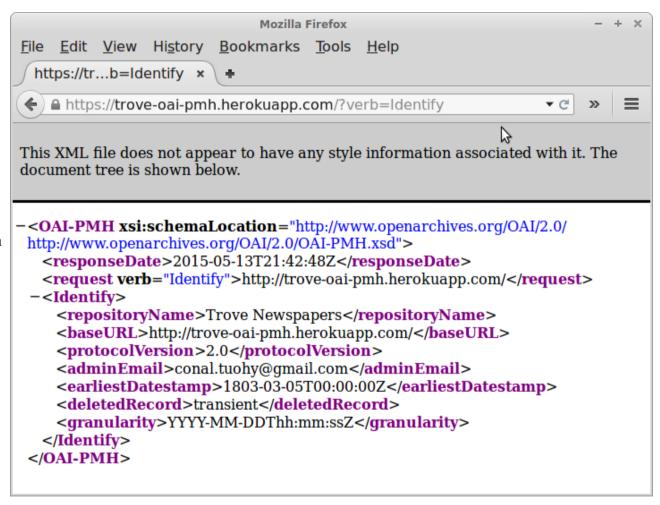
My friend Martin Bush is doing his PhD on popular science in Australia, for which he's researching in Trove's newspaper collection. He is using some software that I wrote to expose Trove's newspapers as full text, using the library data exchange protocol OAI-PMH.

The software (called "Retailer") was needed to convert Trove's own API, which is really a search API, whose focus is on the discovery of individual resources, into an API which is focused on the bulk exchange of metadata.

Martin has used the software to harvest several corpora from Trove:

- 1. Lecture transcripts/detailed reports (about 800 science lectures out of about 3600 lectures overall. Will use "topic modelling" software to analyse this corpus, but even just doing simple stats on eg mention of religion during lectures requires having the corpus at hand rather than searching in Trove.)
- 2. Letters to editors mentioning astronomy (about 1200)
- 3. Ads for the Use of the Globes (about 1500. Will eventually try to then extract names of schools/teachers that appear in the ads)
- 4. Regular columns on science (not done much on yet except compile a list of columns, butt eventually will want to run some stats on numbers of mentions of particular subjects/topics.)

All in the period 1860-1900.



Example 3: PROVisualizer

http://metadata.prov.vic.gov.au/provisualizer/#water

Provisualizer is a visualisation and discovery interface built for the Public Record Office Victoria.

- A dynamic, searchable, network visualisation of the connections between functions and agency in the PROV collection.
- Social-media and embedding tools.
- Links to the related entities in their collection management system.
- Data is licenced as CC-BY.

PROV's goal was to build a visualisation that could be embedded in other websites; to effectively spread bits of PROV's user interface around the web.

This is quite different to PROV's other discovery interfaces; it's built with quite different technology. By opening up their metadata, and publishing it on the web in an open spreadsheet format, they were able to get me to build this interface without having to mess with their internal systems at all.

