



Conal Tuohy

ConalTuohy.com

[@conaltuohy](https://twitter.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 2: Newspaper research

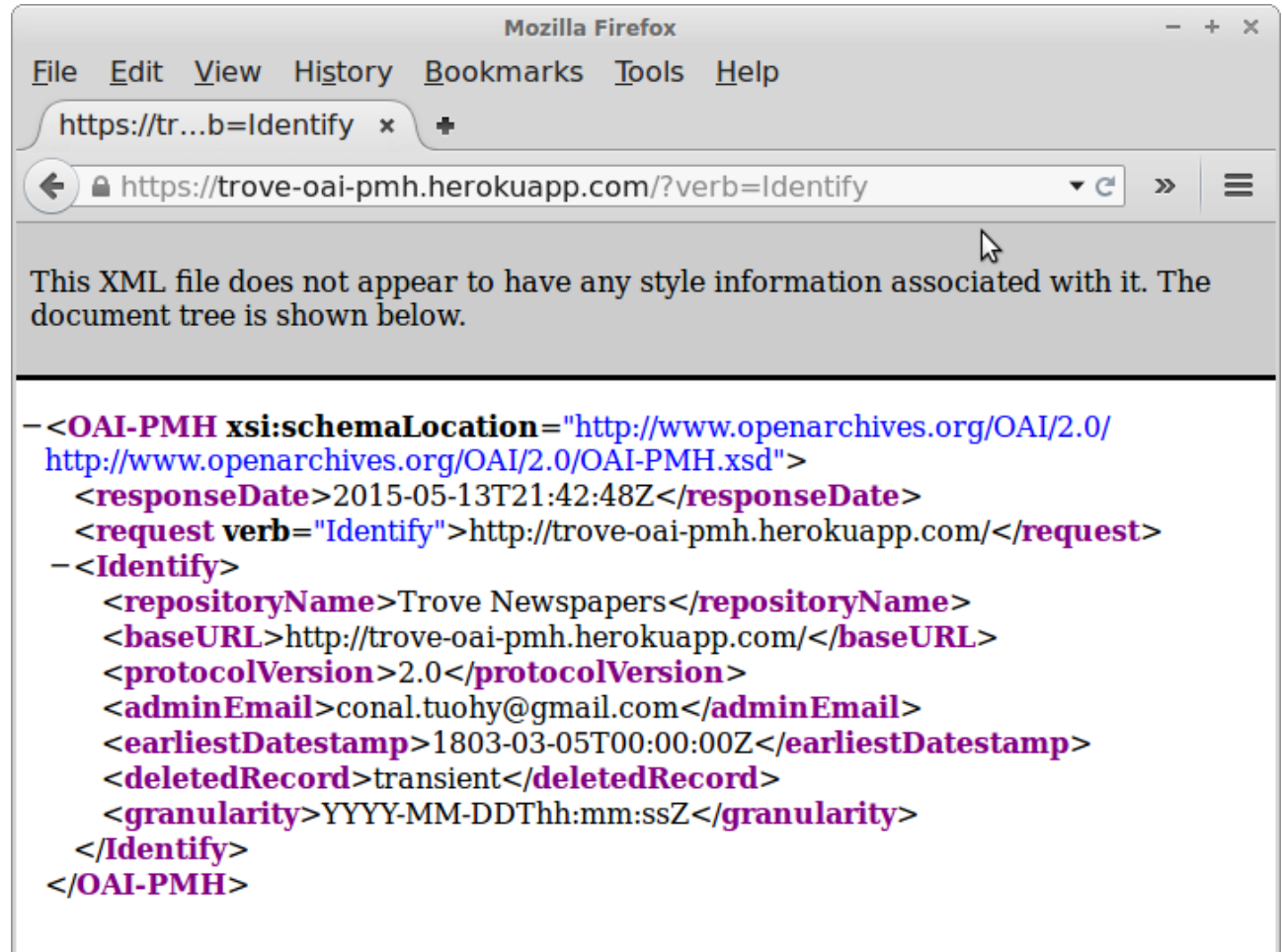
My friend Martin Bush is doing his PhD on popular science in Australia, for which he's researching in Trove's newspaper collection. He is using some software that I wrote to expose Trove's newspapers as full text, using the library data exchange protocol OAI-PMH.

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.



The screenshot shows a Mozilla Firefox browser window with the address bar displaying `https://trove-oai-pmh.herokuapp.com/?verb=Identify`. The page content displays an XML response from the OAI-PMH API. The XML is as follows:

```
-<OAI-PMH xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2015-05-13T21:42:48Z</responseDate>
  <request verb="Identify">http://trove-oai-pmh.herokuapp.com/</request>
  -<Identify>
    <repositoryName>Trove Newspapers</repositoryName>
    <baseURL>http://trove-oai-pmh.herokuapp.com/</baseURL>
    <protocolVersion>2.0</protocolVersion>
    <adminEmail>conal.tuohy@gmail.com</adminEmail>
    <earliestDatestamp>1803-03-05T00:00:00Z</earliestDatestamp>
    <deletedRecord>transient</deletedRecord>
    <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
  </Identify>
</OAI-PMH>
```

