From collection search to collections as data

Tim Sherratt · @wragge · #HOTA2019

These slides:

https://slides.com/wragge/hota-workshop

Refine your results:

Place

- New South Wales (1,161,749)
- Queensland (689,006)
- Victoria (382,009)
- Western Australia (345,487)
- South Australia (310,867)
- ACT (180,429)
- Tasmania (176,568)
- International (42,504)
- National (10,934)
- Northern Territory (10,537)

Title

- The Canberra Times (AC... (179,315)
- The Sydney Morning Her... (100,430)
- The West Australian (P... (97,400)
- The Age (Melbourne, Vi... (93,010)
- The Argus (Melbourne, ... (87,878)

Digitised newspapers and more

Showing: 1 - 20 of at least **3,310,090** Refine search

Sort by: Relevance Sort

RADIO.
The West Australian (Perth, WA : 1879 - 1954) **Tuesday 15 September 1925** p 6 Article

... ***RADIO***. ?Mr. H. Brooghton Jensen, who was recently engaged to make an examination of the ***Radio*** mime ... 68 words

RADIO
The Mail (Adelaide, SA : 1912 - 1954) **Saturday 2 August 1930** p 17 Article Illustrated

... ***RADIO*** Variety will be the keynote of programmes by the Australian Broadcasting Company this week ... 128 words

RADIO.
The Longreach Leader (Qld. : 1923 - 1954) **Friday 26 September 1924** p 14 Article

... licensed: dealers'in all ***radio*** goods, and can quote Sydney prices.,If contemplating installing a set, why ... 111 words

Radio
Queensland Figaro (Brisbane, Qld. : 1901 - 1936) **Saturday 29 October 1927** p 9 Article

... the retail ***Radio*** shops in Brisbane. o Ireland has experienced three eras— The Pagan Era. The Christian ... 458 words

Refine your results:

Place

- New South Wales (1,161,749)
- Queensland (689,006)
- Victoria (382,009)
- Western Australia (345,487)
- South Australia (310,867)
- ACT (180,429)
- Tasmania (176,568)
- International (42,504)
- National (10,934)
- Northern Territory (10,537)

Title

- The Canberra Times (AC... (179,315)
- The Sydney Morning Her... (100,430)
- The West Australian (P... (97,400)
- The Age (Melbourne, Vi... (93,010)
- The Argus (Melbourne, ... (87,878)

Digitised newspapers and more

Showing: 1 - 20 of at least **3,310,090** Refine search

Sort by: Relevance Sort

RADIO.
The West Australian (Perth, WA : 1879 - 1954) **Tuesday 15 September 1925** p 6 Article

... ***RADIO***. ?Mr. H. Brooghton Jensen, who was recently engaged to make an examination of the ***Radio*** mime ... 68 words

RADIO
The Mail (Adelaide, SA : 1912 - 1954) **Saturday 2 August 1930** p 17 Article Illustrated

... ***RADIO*** Variety will be the keynote of programmes by the Australian Broadcasting Company this week ... 128 words

RADIO.
The Longreach Leader (Qld. : 1923 - 1954) **Friday 26 September 1924** p 14 Article

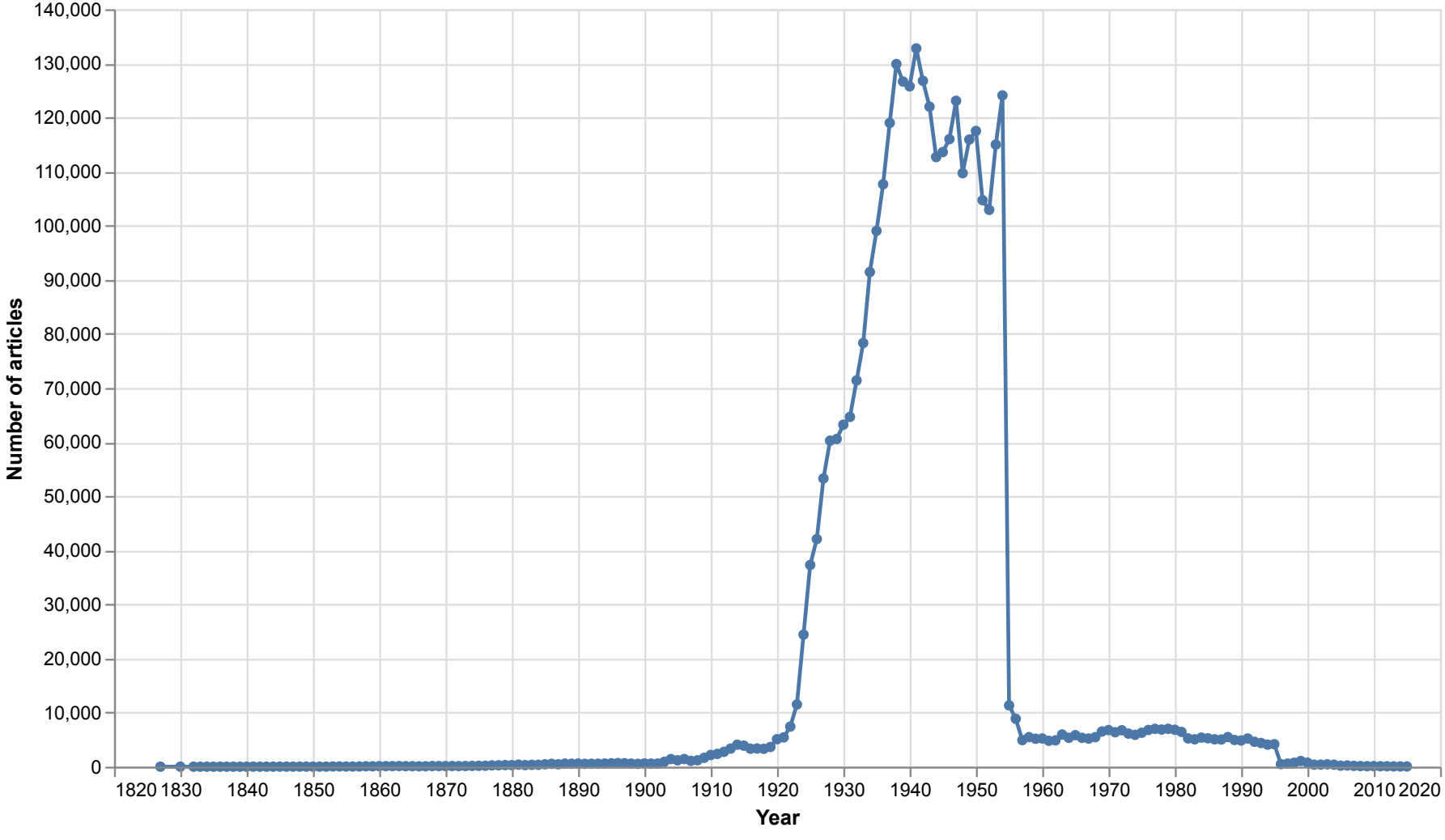
... licensed: dealers'in all ***radio*** goods, and can quote Sydney prices.,If contemplating installing a set, why ... 111 words

Radio
Queensland Figaro (Brisbane, Qld. : 1901 - 1936) **Saturday 29 October 1927** p 9 Article

... the retail ***Radio*** shops in Brisbane. o Ireland has experienced three eras— The Pagan Era. The Christian ... 458 words

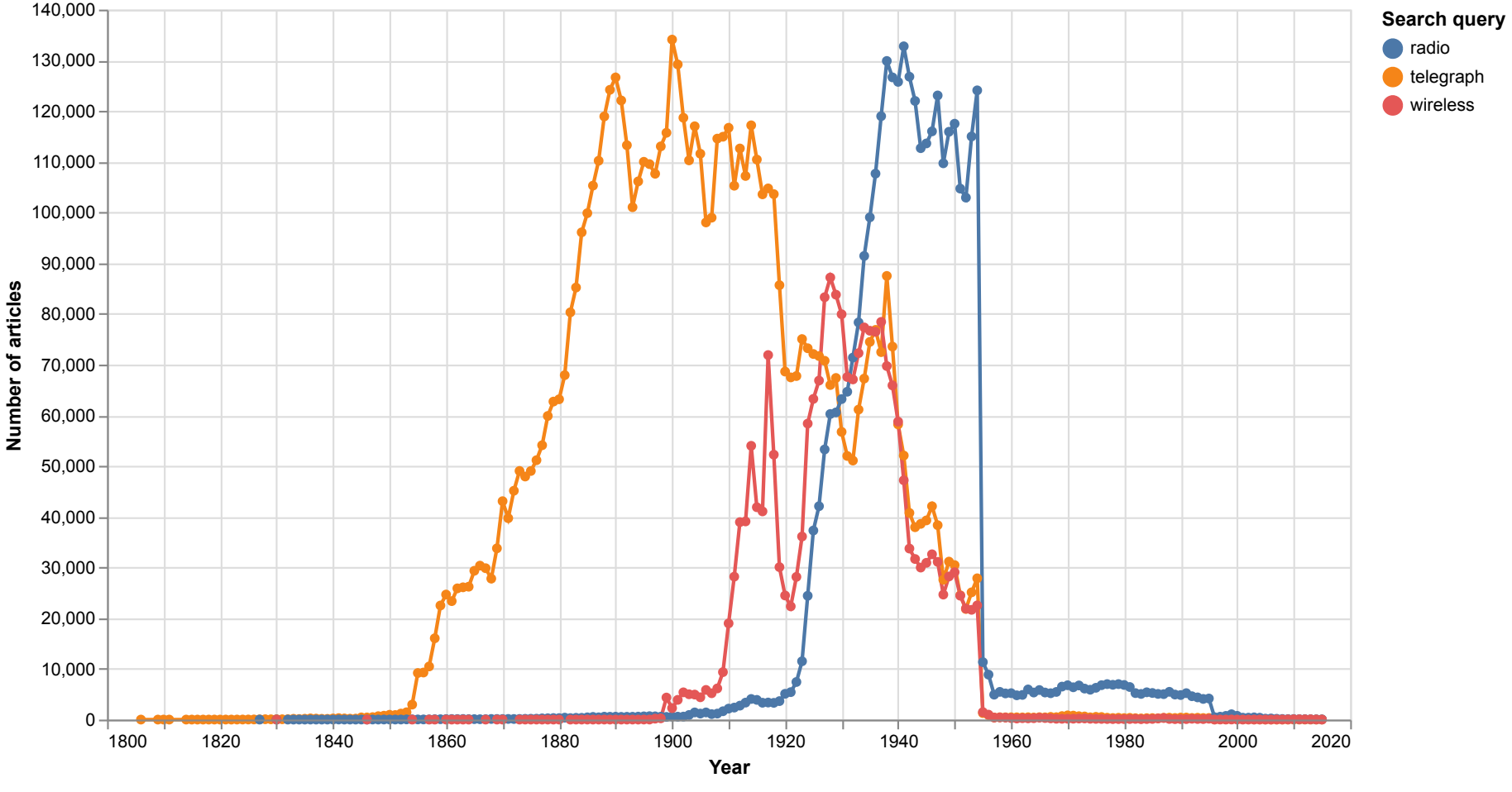
seeing differently

search for 'radio' in Trove newspapers



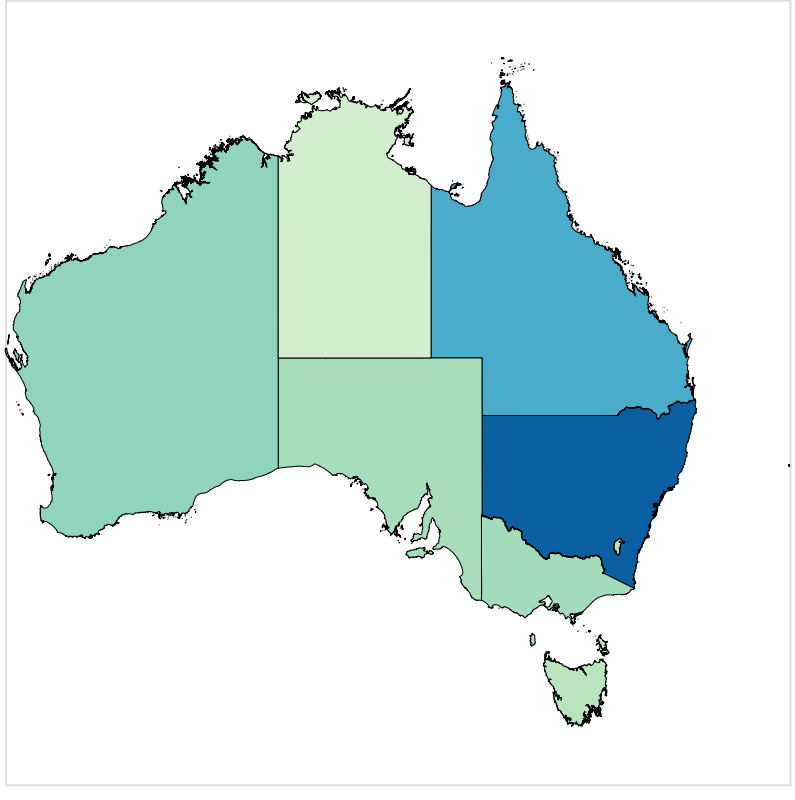
https://glam-workbench.github.io/trove-newspapers/

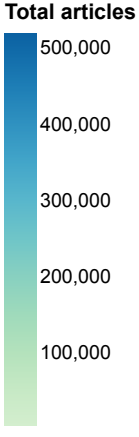
comparing 'radio', 'telegraph', & 'wireless

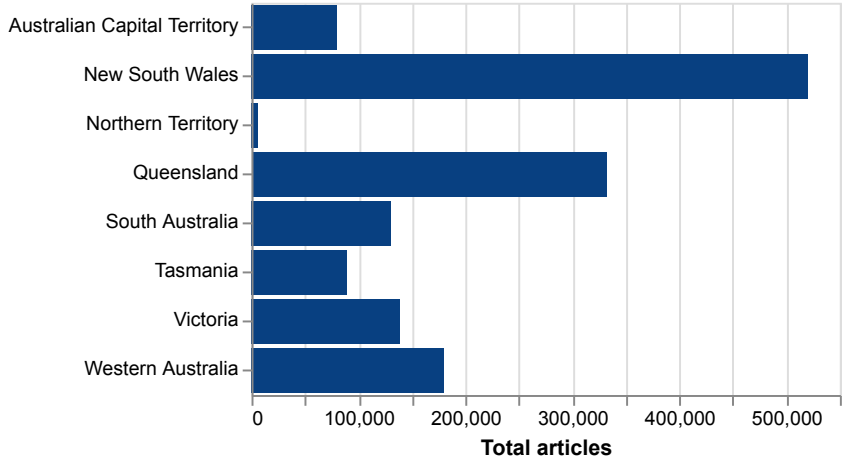


https://glam-workbench.github.io/trove-newspapers/

search for 'radio' showing place of publication

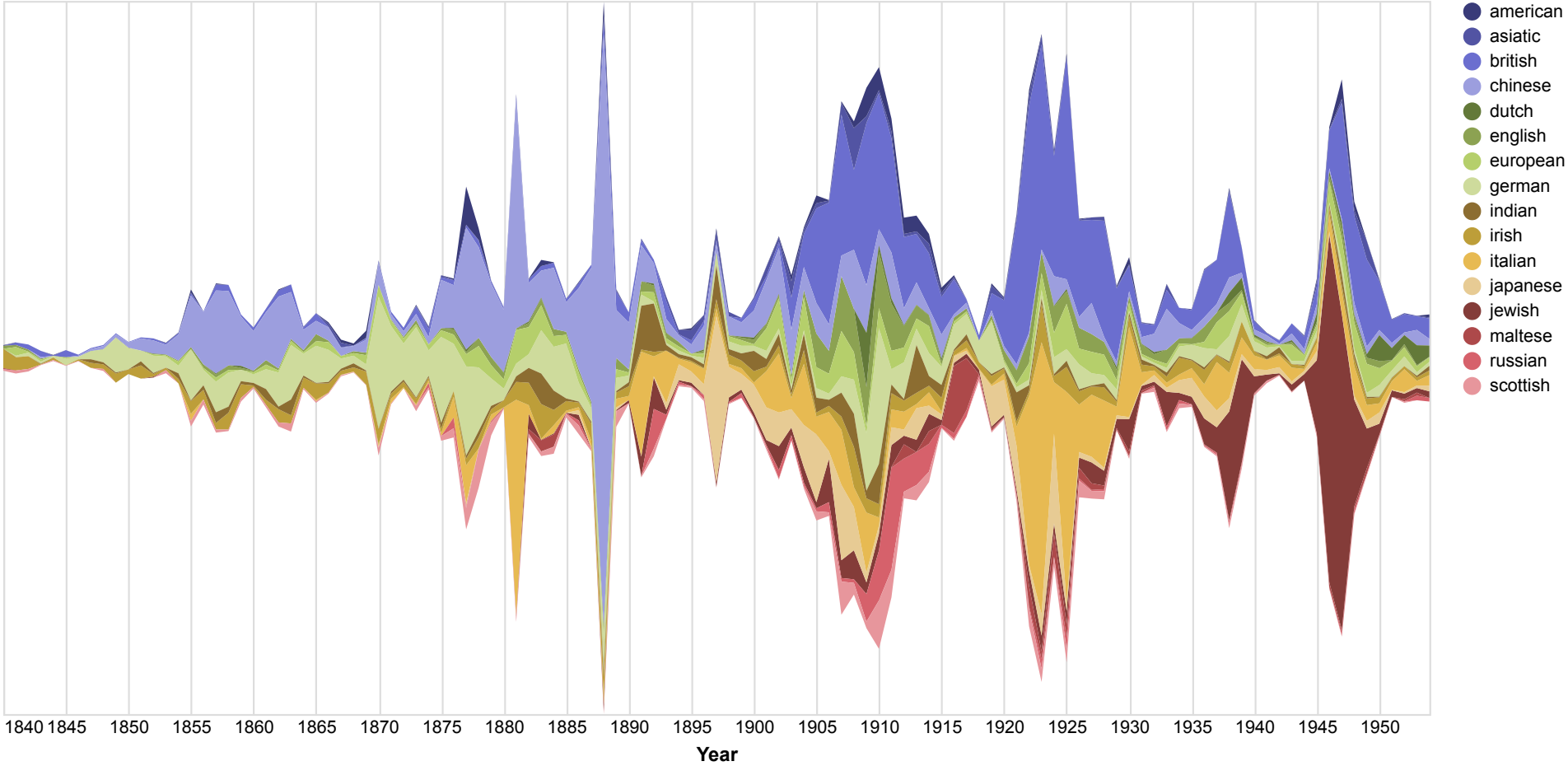






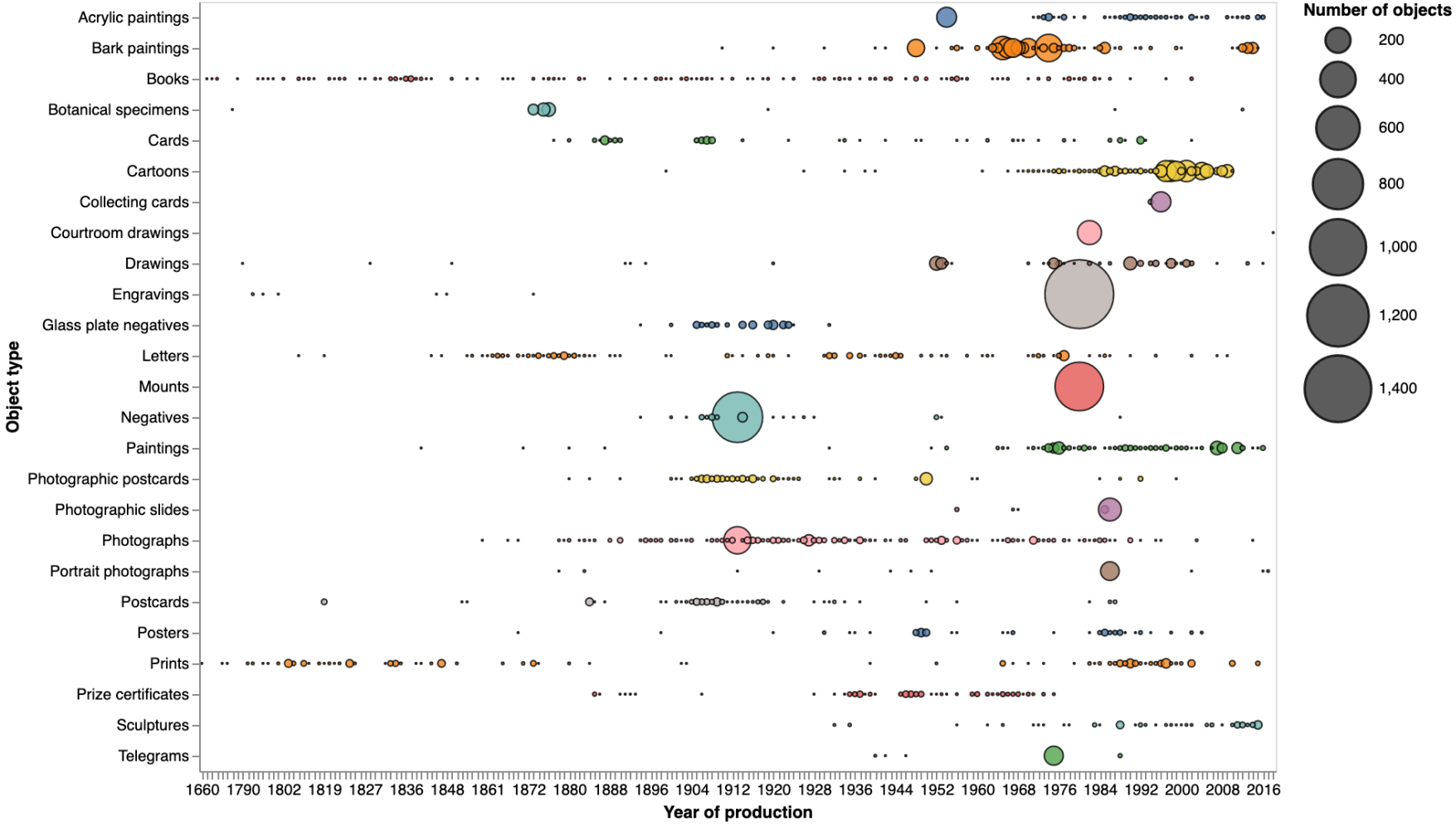
https://glam-workbench.github.io/trove-newspapers/

words from newspapers describing origins of immigrants



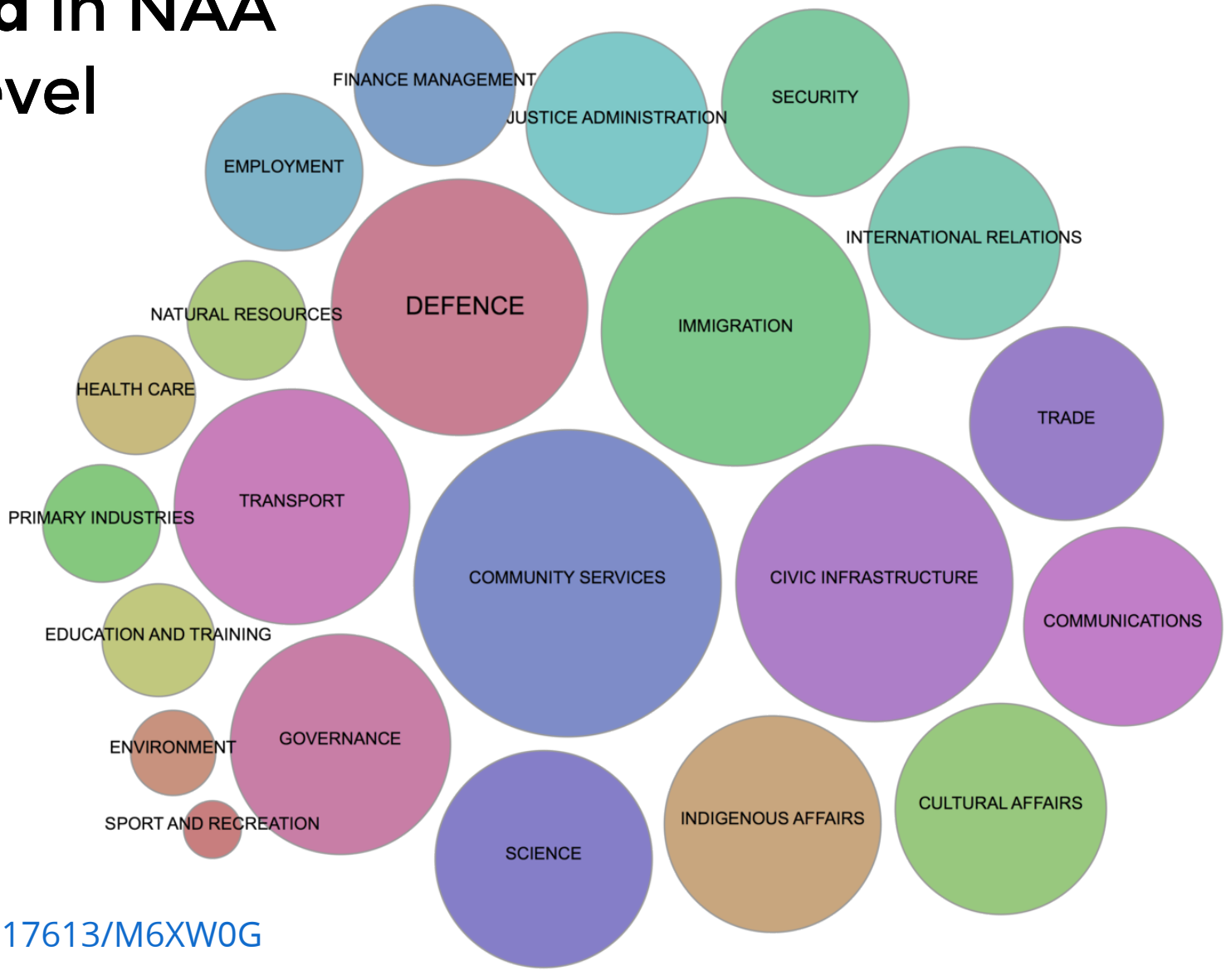
http://timsherratt.org/blog/who-belongs/

objects in the NMA, by year of production



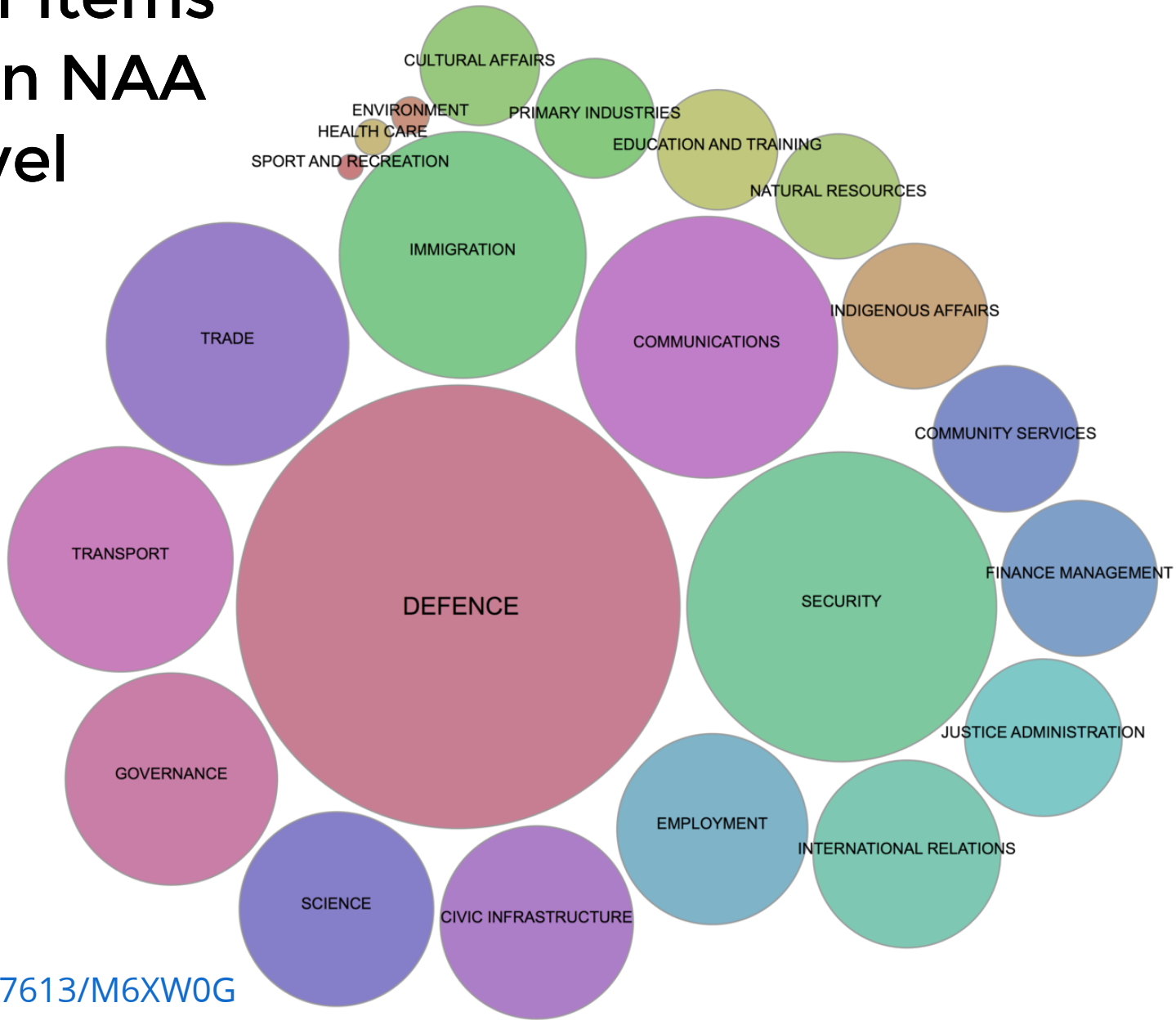
https://glam-workbench.github.io/nma/

number of items **described** in NAA by top-level function



http://dx.doi.org/10.17613/M6XW0G

number of items **digitised** in NAA by top-level function



http://dx.doi.org/10.17613/M6XW0G

Australian aviators in Trove newspapers

https://www.easyzoom.com/imageaccess/9d26953ccdf5475cad9c11f308cd7988

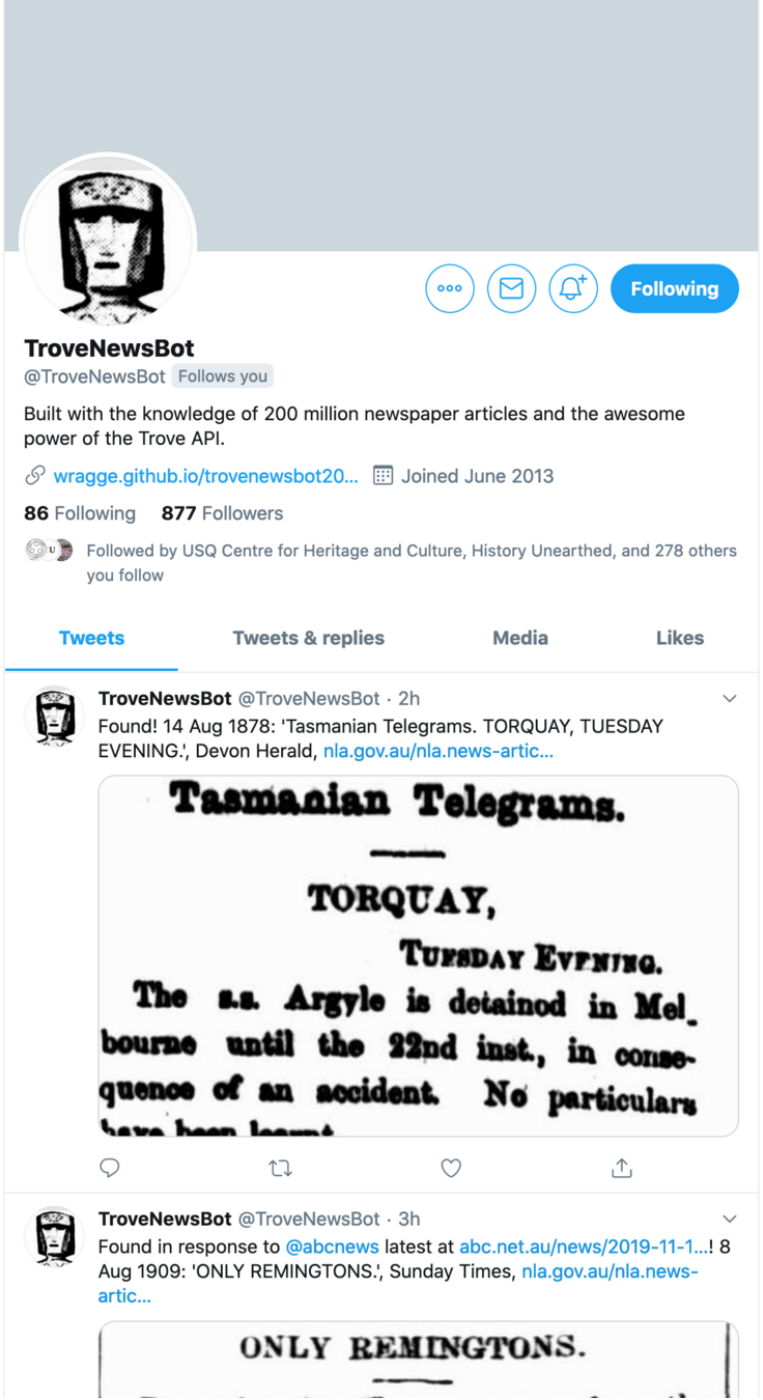
White Australia policy records in the NAA

http://invisibleaustralians.org/faces/

https://www.jstor.org/stable/j.ctvnjbdr0.4

@TroveNewsBot

https://twitter.com/TroveNewsBot



what is GLAM data?

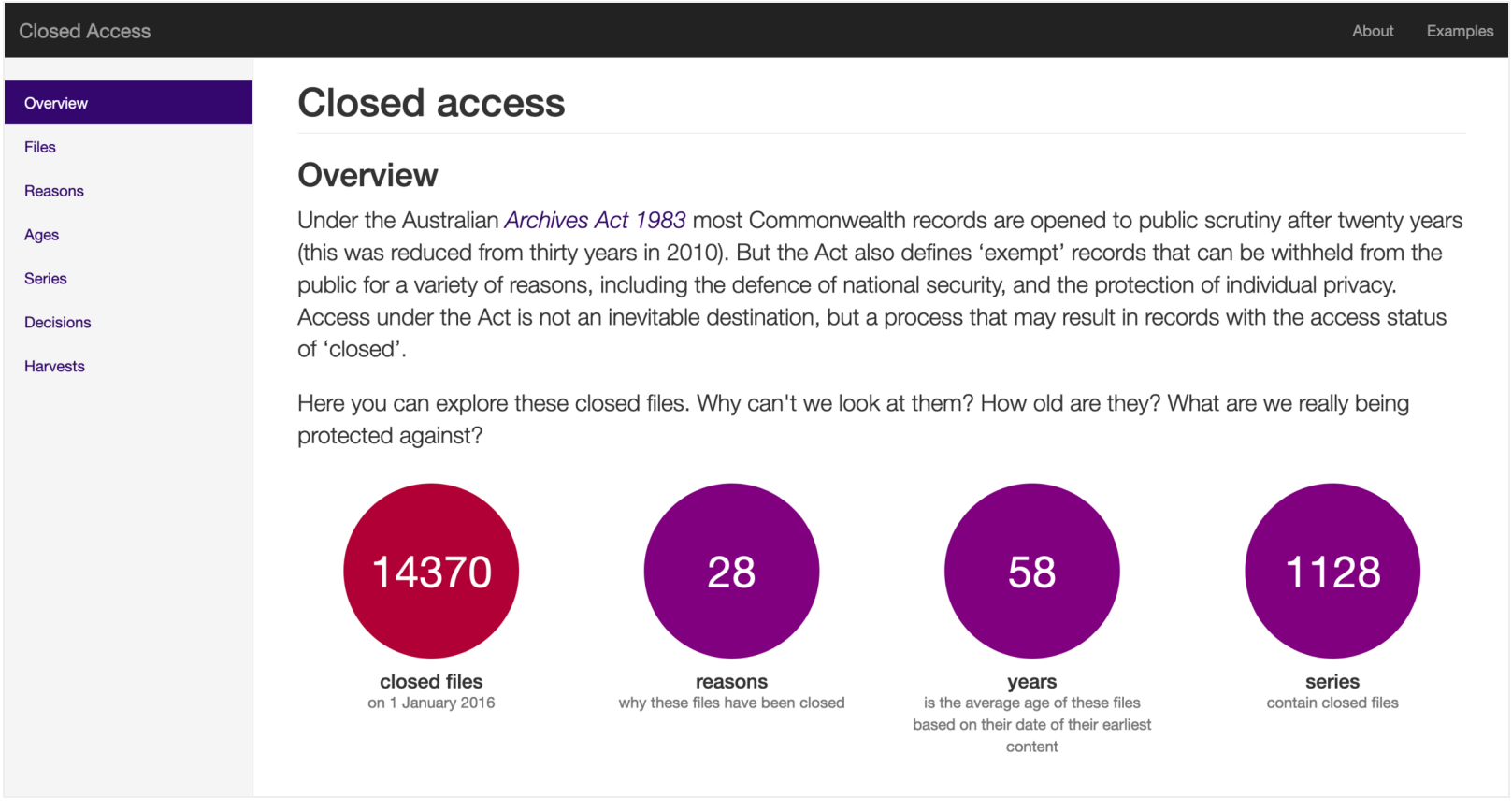
some varieties of GLAM data

- metadata (not content)
- structured text / data
- unstructured text
- images
- derived data
- user generated data
- activity data
- born digital data

metadata

(data about collections)

seeing what we're not allowed to see



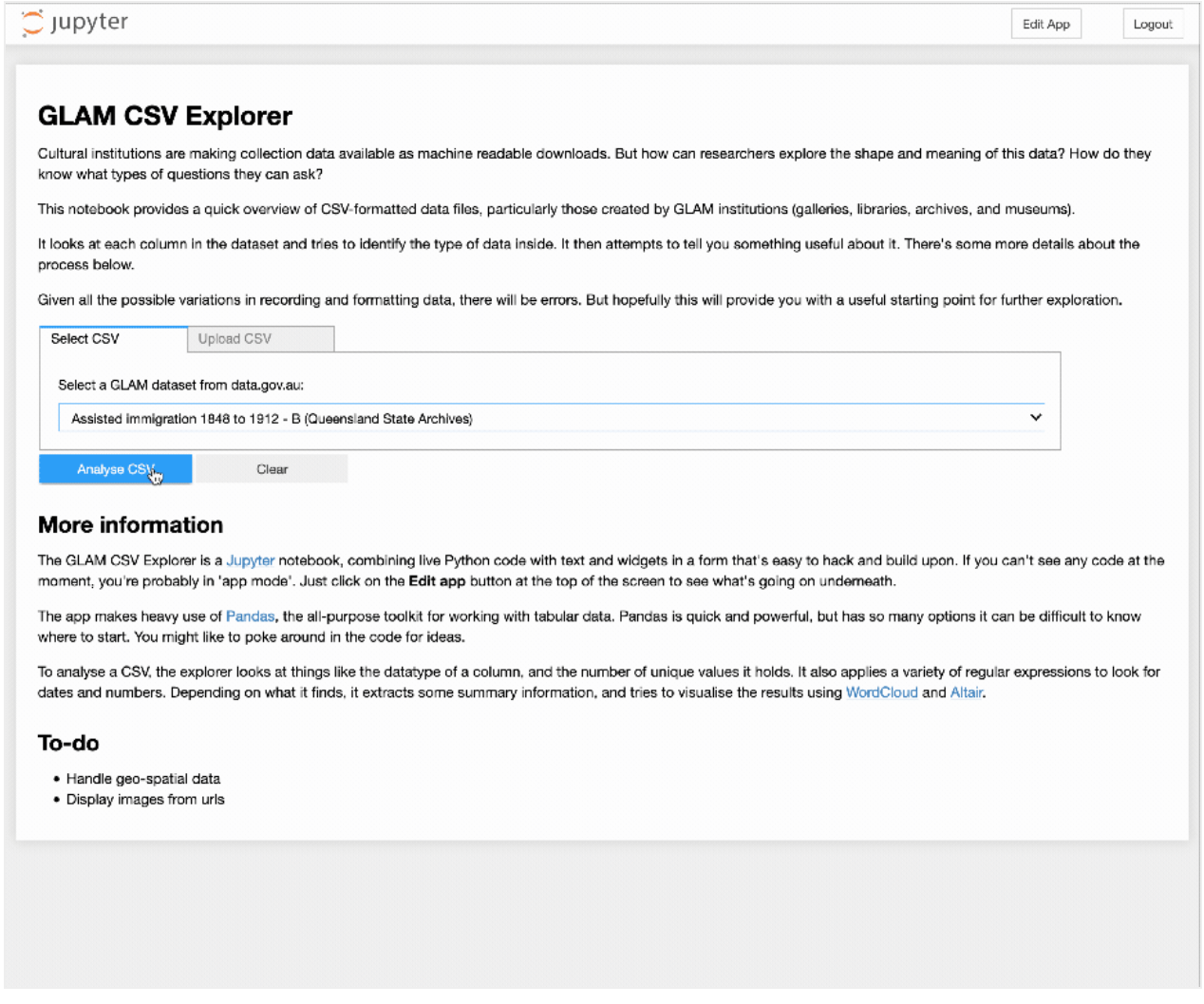
http://closedaccess.herokuapp.com/

http://insidestory.org.au/withheld-pending-advice/

structured data

(think data with rows and columns)

GLAM CSV Explorer



https://glam-workbench.github.io/csv-explorer/

https://glam-workbench.github.io/glam-data-portals/

unstructured text

(lots of words)

text from Trove journals

Digitised journals from Trove with OCRd text

For harvesting details see this notebook, or the digitised journals section of the GLAM Workbench.

This harvest was completed on 27 August 2019.

Number of journals harvested: 720

Number of issues with OCRd text: 33,035

"Coo-ee!" (Bishops Knoll Hospital (Bristol, England))

12 of 13 issues have OCRd text available for download.

- Details on Trove
- Browse issues on Trove
- Download issue data as CSV from CloudStor
- Download all OCRd text from CloudStor

... Review / Remuneration Tribunal

2 of 2 issues have OCRd text available for download.

- Details on Trove
- Browse issues on Trove
- Download issue data as CSV from CloudStor
- Download all OCRd text from CloudStor

https://glam-workbench.github.io/trove-journals/

images

(of images and text)

3,471 Bulletin editorial cartoons

https://glam-workbench.github.io/trove-journals/

derived data

(data you extract from data)

redactions from ASIO files

https://owebrowse.herokuapp.com/redactions/

https://owebrowse.herokuapp.com/redactions/

user-generated data

(data added by the public)

https://github.com/wragge/realface-data

where do you get GLAM data?

Sources of Australian GLAM data

DOI 10.5281/zenodo.3520419

GLAM datasets on data.gov.au

- Human readable list of GLAM datasets harvested from data.gov.au (July 2019)
- CSV formatted list of GLAM datasets harvested from data.gov.au (July 2019)
- CSV formatted list of GLAM datasets (CSVs only) harvested from data.gov.au (July 2019)

Other downloadable datasets

Full text

- Commonwealth Parliamentary Debates (Hansard), 1901-1980 (harvested from Parliamentary Library)
- Hansard interjections
- Australian Government Gazettes (1832-1968) (Trove)
- Federal Election speeches (Museum of Australian Democracy)
- Prime Ministers transcripts (harvested from DPMC)
- OCRd text from Trove digitised books (and ephemera) (harvested from Trove)
- OCRd text from the Internet Archive of 'Australian' books listed in Trove (harvested from Trove and Internet Archive)
- OCRd text of Trove digitised journals (harvested from Trove)
- Parliamentary press releases relating to immigrants and refugees (harvested from Trove and Parliamentary Library)
- Real Face of White Australia data (transcribed from National Archives of Australia: ST84/1)

https://glam-workbench.github.io/glam-data-list/

but if it's not already packaged for download, you can...

- harvest data from APIs...
- extract data from web pages by screen scraping...

it all seems too hard!

the GLAM Workbench is here to help!

GLAM Workbench

Search

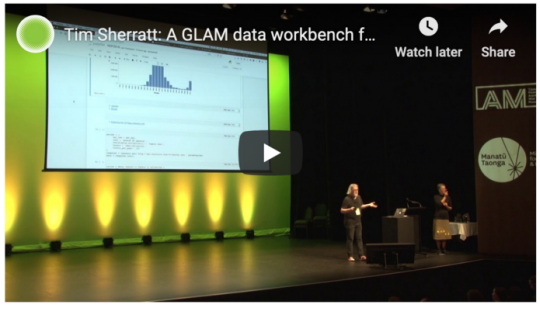
GLAM-Workbench
29 Repositories

GLAM Workbench

- Home
- Some background
- Suggest a topic
- Getting started
- General
- Trove
- National Archives of Australia
- State Library of NSW
- NSW State Archives
- Queensland State Archives
- Australian government
- National Archives of NZ
- DigitalNZ
- Te Papa
- Library Archives Canada

Welcome to the wonderful world of GLAM data!

Here you'll find a collection of tools and examples to help you work with data from galleries, libraries, archives, and museums (the GLAM sector), focusing on Australia and New Zealand.



What is GLAM data?

When we talk about GLAM data we're usually referring to the collections held by cultural institutions – books, manuscripts, photographs, objects, and much more. We're used to exploring these collections through online search interfaces or finding aids, but sometimes we want to do more – instead of a list of search results on a web page, we want access to the underlying collection data for analysis, enrichment, or visualisation. We want collections as data.

This GLAM Workbench shows you how to create your own research datasets from a variety of GLAM collections. In some cases cultural institutions provide direct access to collection data through APIs (Application Programming Interfaces) or data downloads. In other cases we have to find ways of extracting data from web interfaces – a process known as screen-scraping. Here you'll find examples of all these approaches, as well as links to a number of pre-harvested datasets.

Table of contents

https://glam-workbench.github.io/

Jupyter notebooks?

Using Jupyter notebooks in the GLAM Workbench

The GLAM Workbench includes many Jupyter notebooks. Jupyter lets you combine text, images, and live code within a single web page. So not only can you read about collections data, you can download it, analyse it, and visualise it – all within your browser!

While the notebooks often include some fairly intimidating looking code, you don't need to understand the code to use them. As explained below, there's just a couple of basic conventions you need to keep in mind when running Jupyter notebooks. Once you've mastered these, you'll be able to use any of the tools or examples in this workbench.

Of course, once you've developed a bit of confidence, you might want to start playing around with the code. That's how you learn. The GLAM Workbench isn't just a collection of tools, it's a starting point – from here you can explore, extend, and experiment!

Running code in a notebook

Most of the notebooks in the GLAM Workbench include snippets of real code. You can use this code to do things like download data, or create charts. The programming language used here is Python. It's popular in the data sciences and is generally pretty easy for humans to understand.

The code in Jupyter notebooks is contained in cells, or boxes, on the page – you can identify code cells by the borders around them.

To run code snippets:

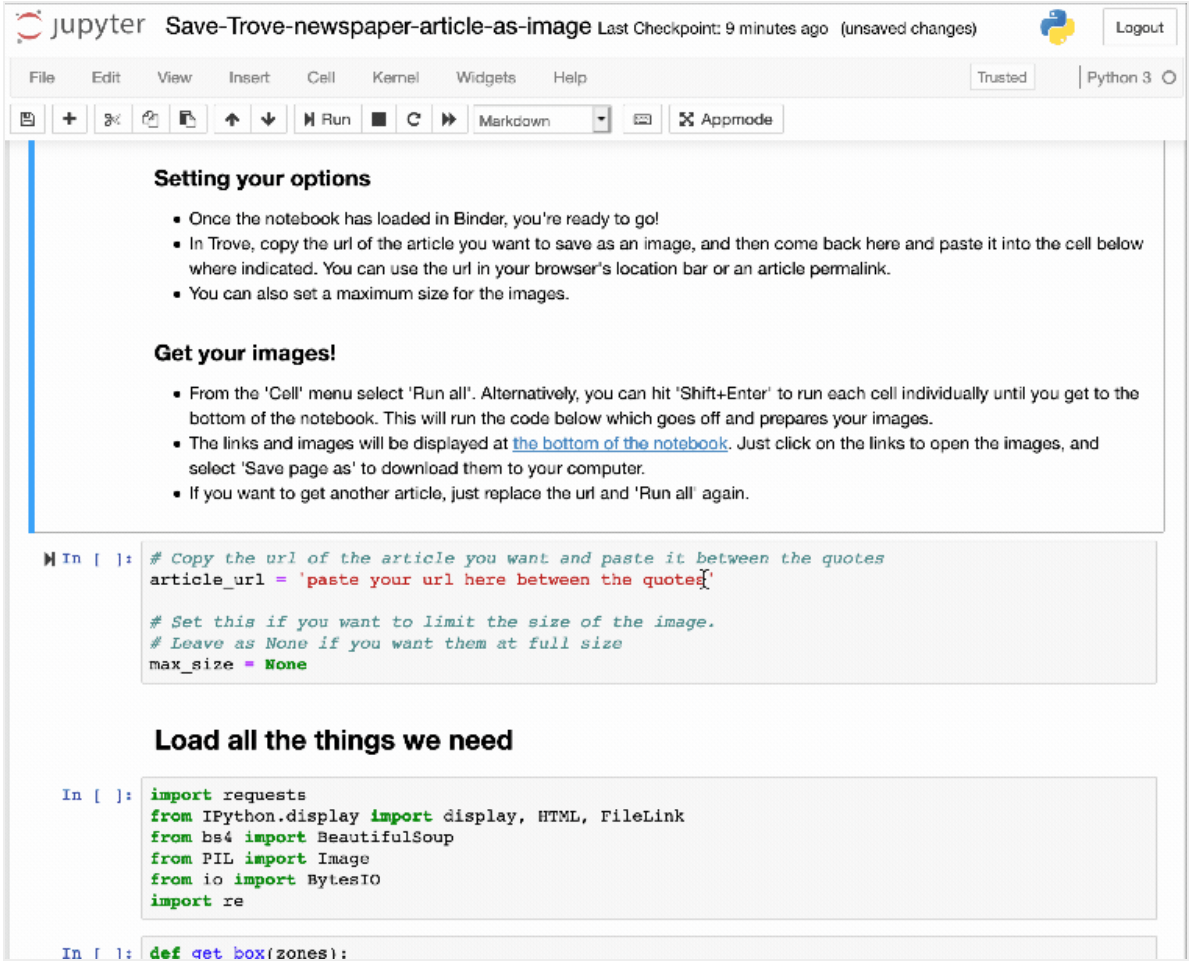
1. Click on the code cell (you'll see the cell becomes highlighted).
2. Hit **Shift+Enter** (the code will run and you'll be moved on to the next cell).

That's it – try it with the cell below!

In [ ]:
```
# CLICK ON ME AND THEN HIT SHIFT+ENTER!

# This makes the datetime module available to use
import datetime

# This creates a variable called 'date_now' and uses the datetime.date.today() function to set it to today's date.
date_now = datetime.date.today()

# This displays a nicely-formatted string containing the date
print(f'Congratulations! You ran the code in this cell on {date_now}')

# Hey! Have you noticed that lines starting with '#' are comments? They can help you understand what's going on in the
```

Try it live!

notebooks can be tools or hacks...



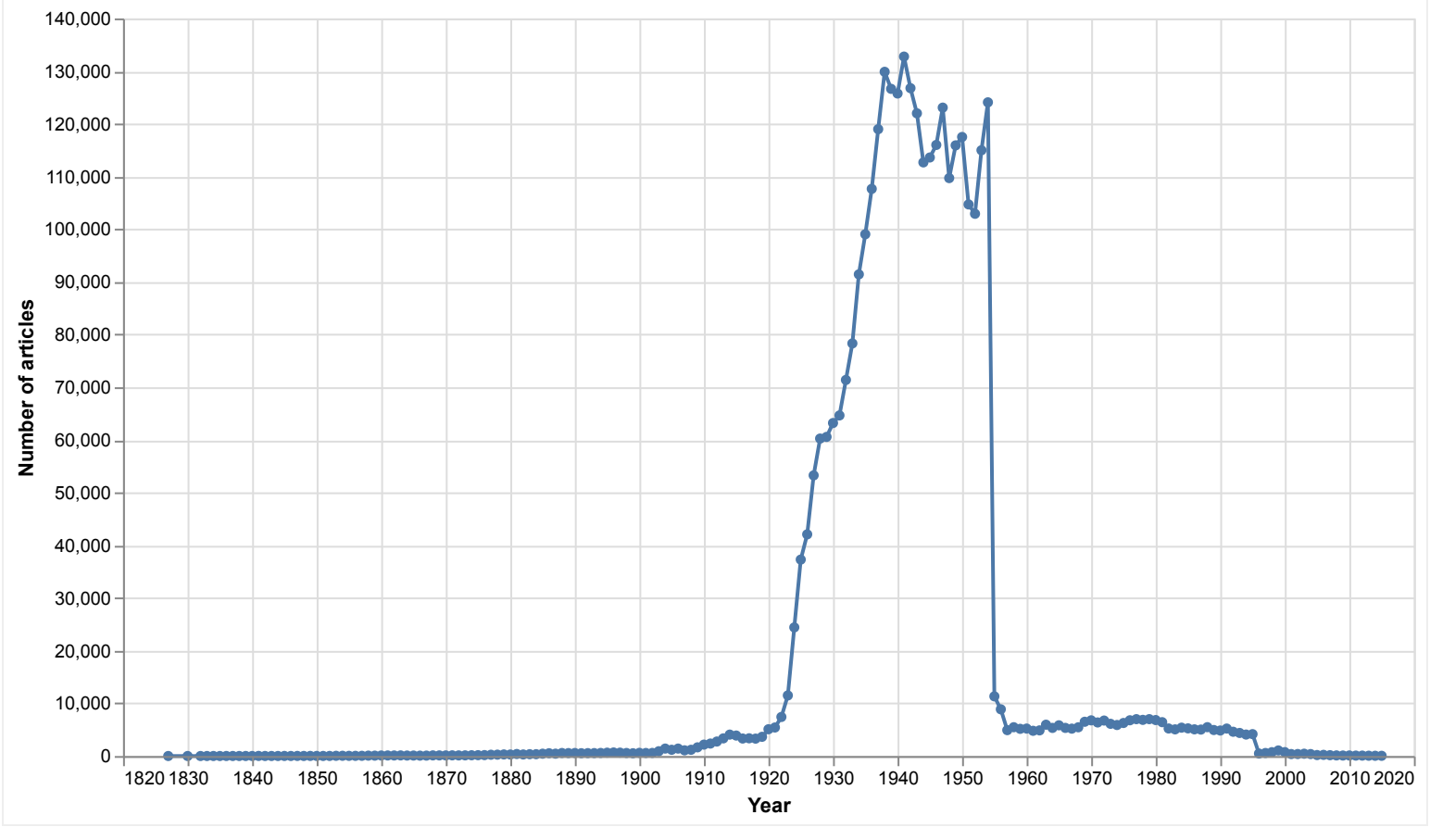
Try it live!

...or even simple apps

- Create a thumbnail image from a Trove newspaper article
- Save a Trove newspaper page as a (high-res) image
- Download the contents of a digitised file from RecordSearch

asking questions of data

what happened to radio in 1955?



https://glam-workbench.github.io/trove-newspapers/

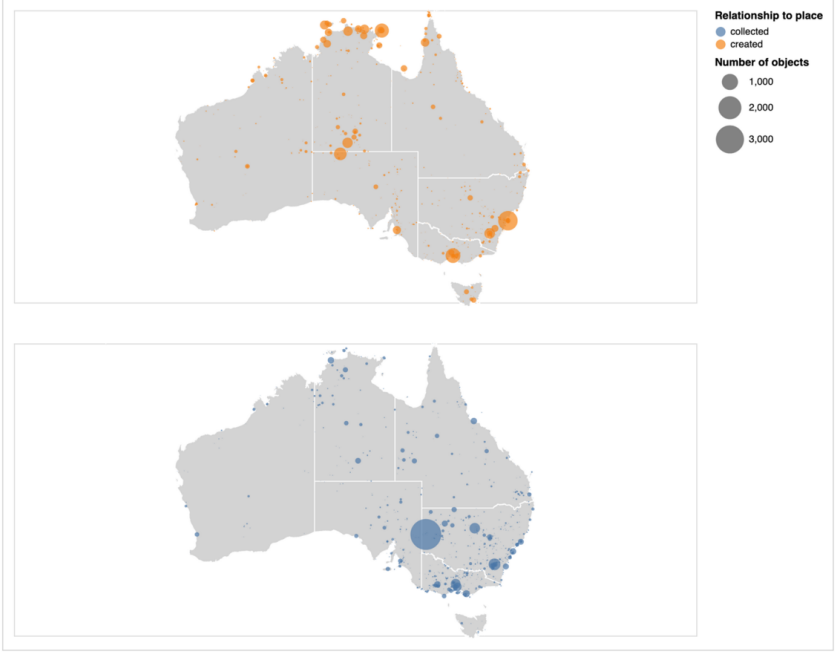
exploring data

what can I do with the NMA API?

Explore places associated with collection objects

In this notebook we'll explore the spatial dimensions of the object data. Where were objects created or collected? To do that we'll extract the nested spatial data, see what's there, and create a few maps.

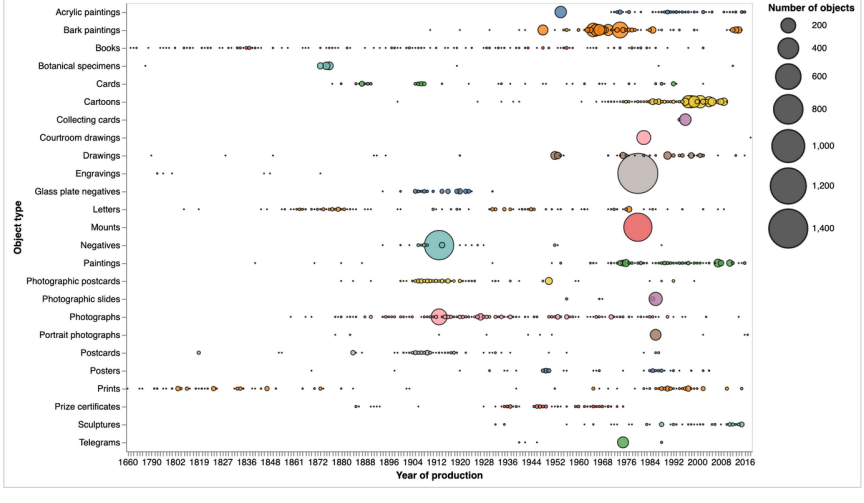
- Download from GitHub
- View using NBViewer
- Run live on Binder



Explore collection objects over time

In this notebook we'll explore the temporal dimensions of the object data. When were objects created, collected, or used? To do that we'll extract the nested temporal data, see what's there, and create a few charts.

- Download from GitHub
- View using NBViewer
- Run live on Binder



https://glam-workbench.github.io/nma/

playing with data

what happens when you view interjections as tweets?

http://nla.gov.au/nla.obj-162853701

Real Words Imagined Tweets

How has social media changed political discourse? Let's flip the question around and look at past political discussions as if they were being shared through Twitter. What changes?

Here you can explore interjections documented in Hansard, the official record of Australia's parliamentary proceedings, between 1901 and 1980. The words, speakers, and dates are real. The Twitter handles, activity stats, and times are invented.

View 3 new Tweets

Hattil Foll
@SenatorFoll

Only about four months of the financial year remain. historichansard.net/sena

12:13 AM - 5 Feb 1926

0 7 51

John Carrick
@SenatorCarrick

The Senate can judge why the Opposition does not want the document incorporated in Hansard. I now table it. historichansard.net/sena

8:15 PM - 16 Sep 1980

1 23 48

Filter

Explore

- beer OR money
- "White Australia"
- refugee
- rabbits
- god
- king OR queen
- or just wait a few seconds…

The data

Over 900,000 interjections were extracted from the Hansard XML repository, which was harvested from the ParlInfo database published on the Parliament of Australian website.

Thanks to @legostormtroopr and the Psephos site for the avatars.

In keeping with Twitter conventions, the interjections displayed here are less than 116 characters long – allowing room for a link to the full debate during

http://hansard-interjections.herokuapp.com/tweets/

http://timsherratt.org/blog/muliplication-of-contexts/

what becomes possible
when I...?

one possible pathway...

1. visualise searches in Trove newspapers
2. find patterns, ask questions
3. zoom in on points of interest
4. harvest article text and metadata
5. explore harvested data in detail

more resources...

- Trove tips & tricks
- Trove as a platform for digital research
- Digital tools and such like
- GLAM collections as data (video)

https://timsherratt.org/