hacking heritage

exploring the limits of access

Tim Sherratt · @wragge · timsherratt.org

these slides

https://slides.com/wragge/hacking-heritage-openglamnow

> A hack can be elegant or kludgy, authored from scratch or patched together and remixed—the important thing is getting things done, pushing the boundaries of what the humanities can do, what effects it can have in the world, and where.

— *Mark J Olson, 'Hacking the humanities: Twenty-first-Century literacies and the 'becoming other' of the humanities'*

access is constructed

Explore Trove's digitised journals

112 journals and 86,211 articles added since 4 July 2019

There's lots of exciting new digitised content being added to Trove's journals zone, but it's not always easy to find and search.

This page lists journals that have been digitised by the NLA and have searchable records for individual articles. This means you can search *inside* the journal, just like you do in the newspapers zone.

To search inside these journals, just click on the titles you're interested in below, then enter keywords in the search box.

Enter your search keywords

No titles selected.

Search Trove

Clear everything

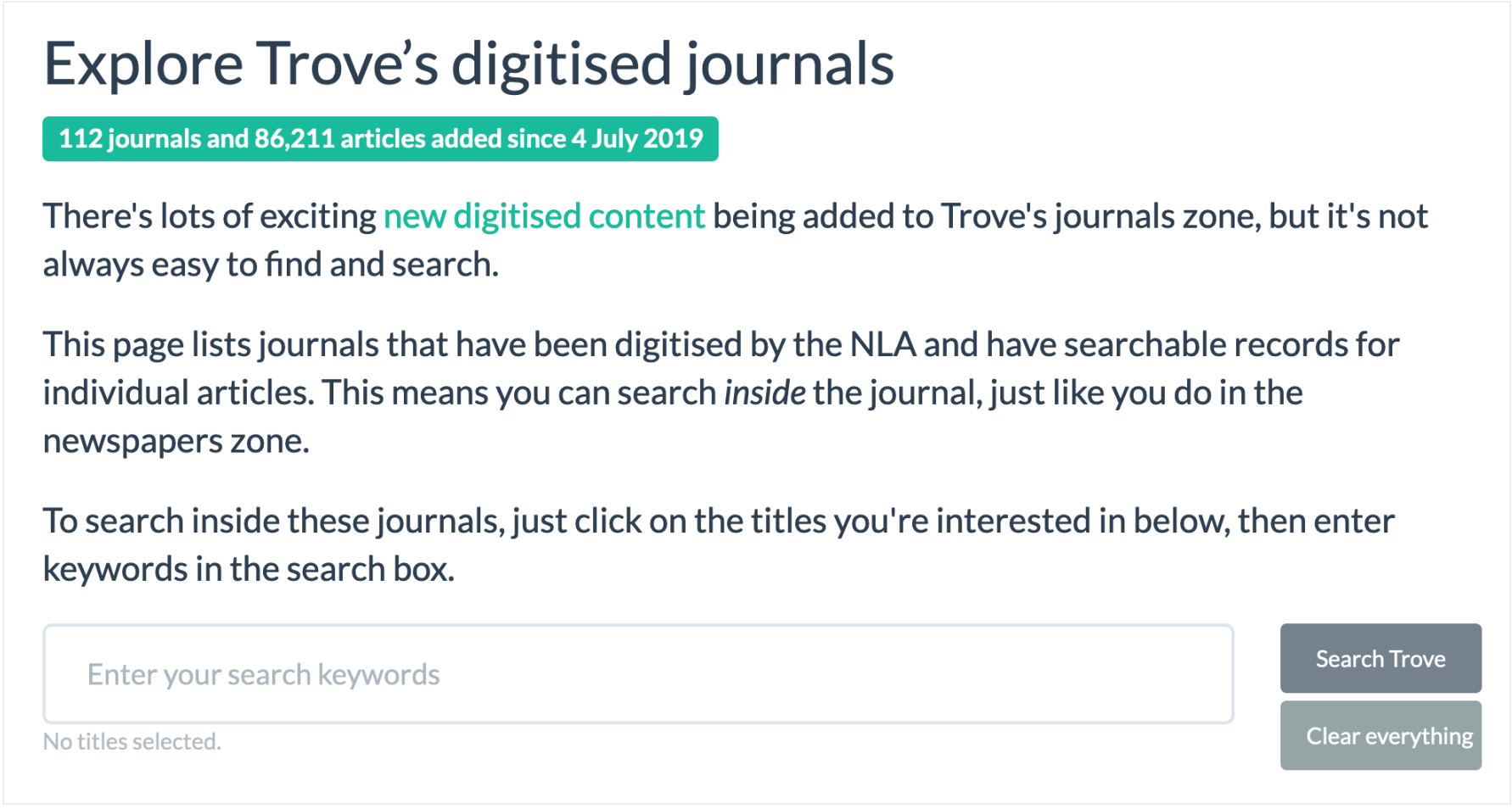
https://trove-titles.herokuapp.com/

- what's available?
- is it in the form I need?
- do I have the technical skills?
- what does it really mean?

what's available?

I don't know where to start...

an alternative interface



https://trove-titles.herokuapp.com/

some random hacks

Get a random illustrated advertisement from the *Australian Womens Weekly*

In [9]:
```
get_random_article(title='112', illustrated='true', category='Advertising')
```

Out[9]:
```
{'id': '53250867',
 'url': '/newspaper/53250867',
 'heading': 'Advertising',
 'category': 'Advertising',
 'title': {'id': '112',
  'value': "The Australian Women's Weekly (1933 - 1982)"},
 'date': '1980-06-18',
 'page': 101,
 'pageSequence': 101,
 'relevance': {'score': '0.24747185', 'value': 'may have relevance'},
 'troveUrl': 'https://trove.nla.gov.au/ndp/del/article/53250867?searchTerm=%22had%22'}
```

Get a random cartoon

In [10]:
```
get_random_article(illtype='Cartoon')
```

Out[10]:
```
{'id': '75198942',
 'url': '/newspaper/75198942',
 'heading': "Don't Tell Auntie An Hilarious Christmas-time Episode",
 'category': 'Article',
 'title': {'id': '251', 'value': 'Sunshine Advocate (Vic. : 1924 - 1954)'},
 'date': '1937-12-17',
 'page': 1,
 'pageSequence': '1 S',
 'relevance': {'score': '1.9126042', 'value': 'likely to be relevant'},
 'snippet': "George Cunningham was not carrying his auntie's luggage; he was only directing the efforts of all the
 'troveUrl': 'https://trove.nla.gov.au/ndp/del/article/75198942?searchTerm=%22hadn%22'}
```

Get a random article from 1930

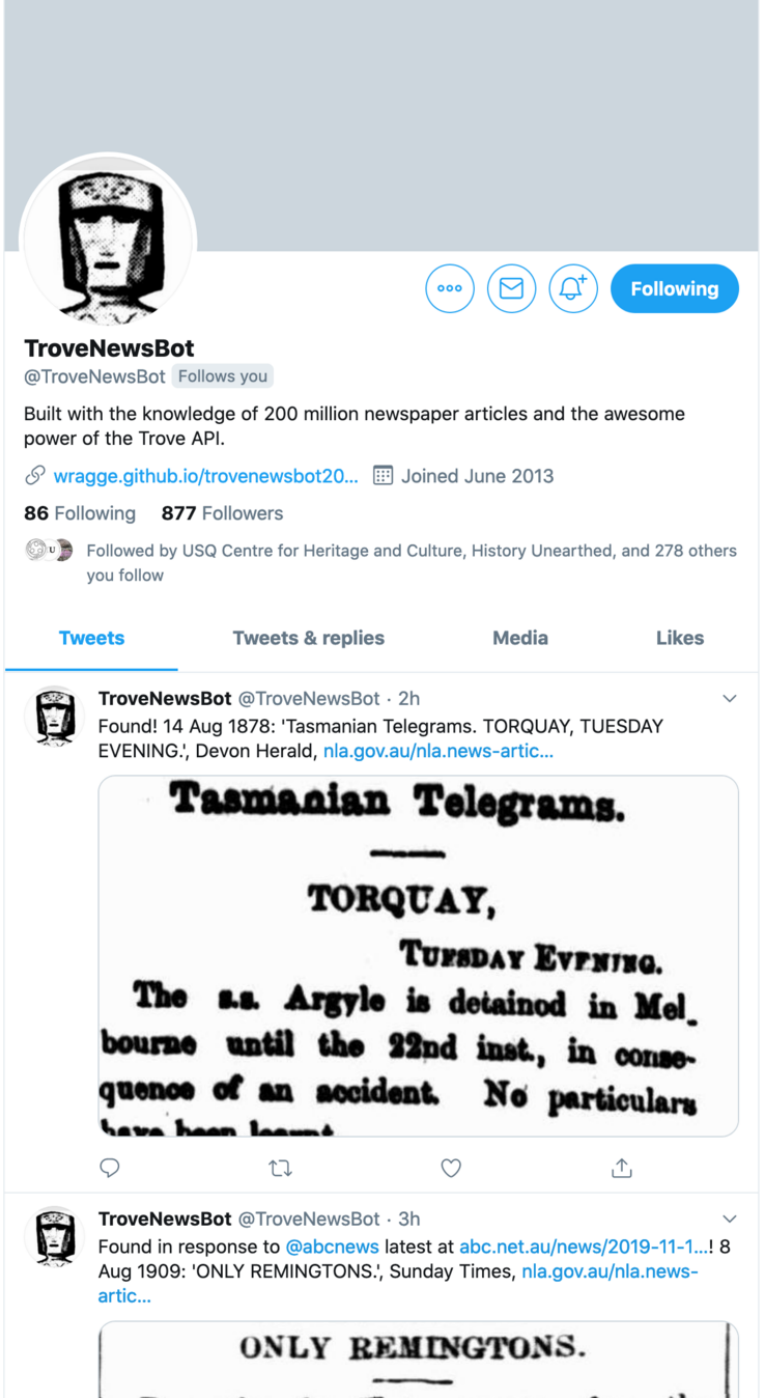
In [11]:
```
get_random_article(year='1930')
```

Out[11]:
```
{'id': '113686939',
 'url': '/newspaper/113686939',
 'heading': 'DID BUILDER ERR? President Townsend Outspoken',
 'category': 'Article',
 'title': {'id': '410',
  'value': 'Gilgandra Weekly and Castlereagh (NSW : 1929 - 1942)'},
 'date': '1930-11-20',
 'page': 7,
 'pageSequence': 7,
 'relevance': {'score': '0.9030686', 'value': 'likely to be relevant'},
 'snippet': '"In this case it appears that the Council has been deliberately flouted," declared the President (Cr.
 'troveUrl': 'https://trove.nla.gov.au/ndp/del/article/113686939?searchTerm=%22not%22'}
```

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

@TroveNewsBot

https://twitter.com/TroveNewsBot



lists can be useful too!

GLAM datasets harvested from data.gov.au (8 July 2019)

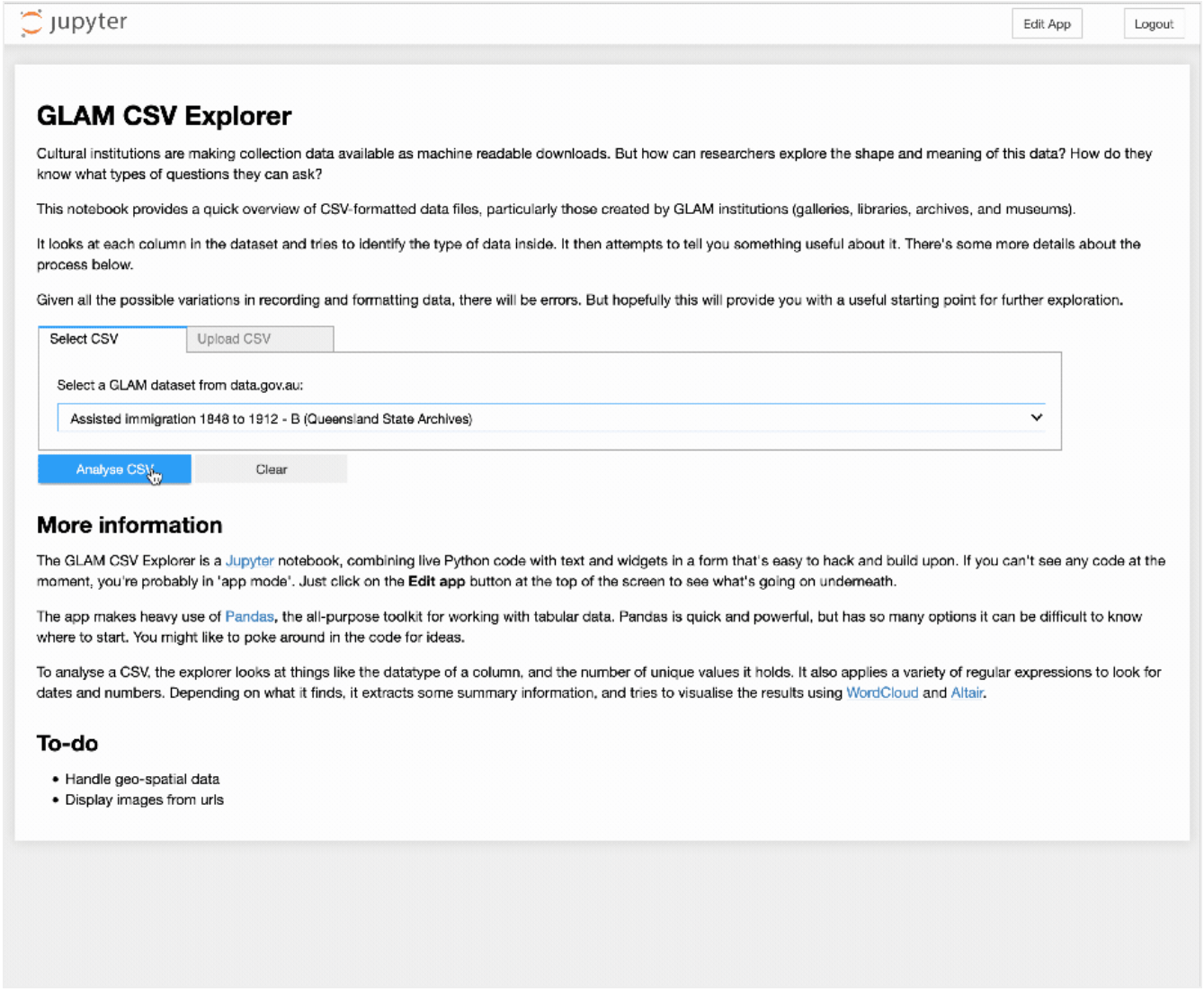
See GLAM Workbench for harvesting code and summary.

- Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS)
- Australian Museum
- History Trust of South Australia
- Libraries Tasmania
- Mount Gambier Library
- Museum of Applied Arts and Sciences
- NSW State Archives
- National Archives of Australia
- National Library of Australia
- National Portrait Gallery
- PROV Public Record Office
- Queensland State Archives
- South Australian Museum
- State Library of NSW
- State Library of New South Wales
- State Library of Queensland
- State Library of South Australia
- State Library of Victoria
- State Library of Western Australia
- State Records
- State Records Office of Western Australia
- Tasmanian Museum and Art Gallery
- Western Australian Museum

```
Creative Commons Attribution                          250
Creative Commons Attribution 3.0 Australia            244
Creative Commons Attribution 4.0                      237
Creative Commons Attribution 4.0 International        146
Creative Commons Attribution 2.5 Australia             32
Creative Commons Attribution-NonCommercial             10
Other (Open)                                            5
notspecified                                            5
Creative Commons Attribution Share-Alike 4.0            3
Creative Commons Attribution 3.0                        3
Creative Commons Attribution Non-Commercial 4.0         2
Custom (Other)                                          1
```

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

but what's inside?

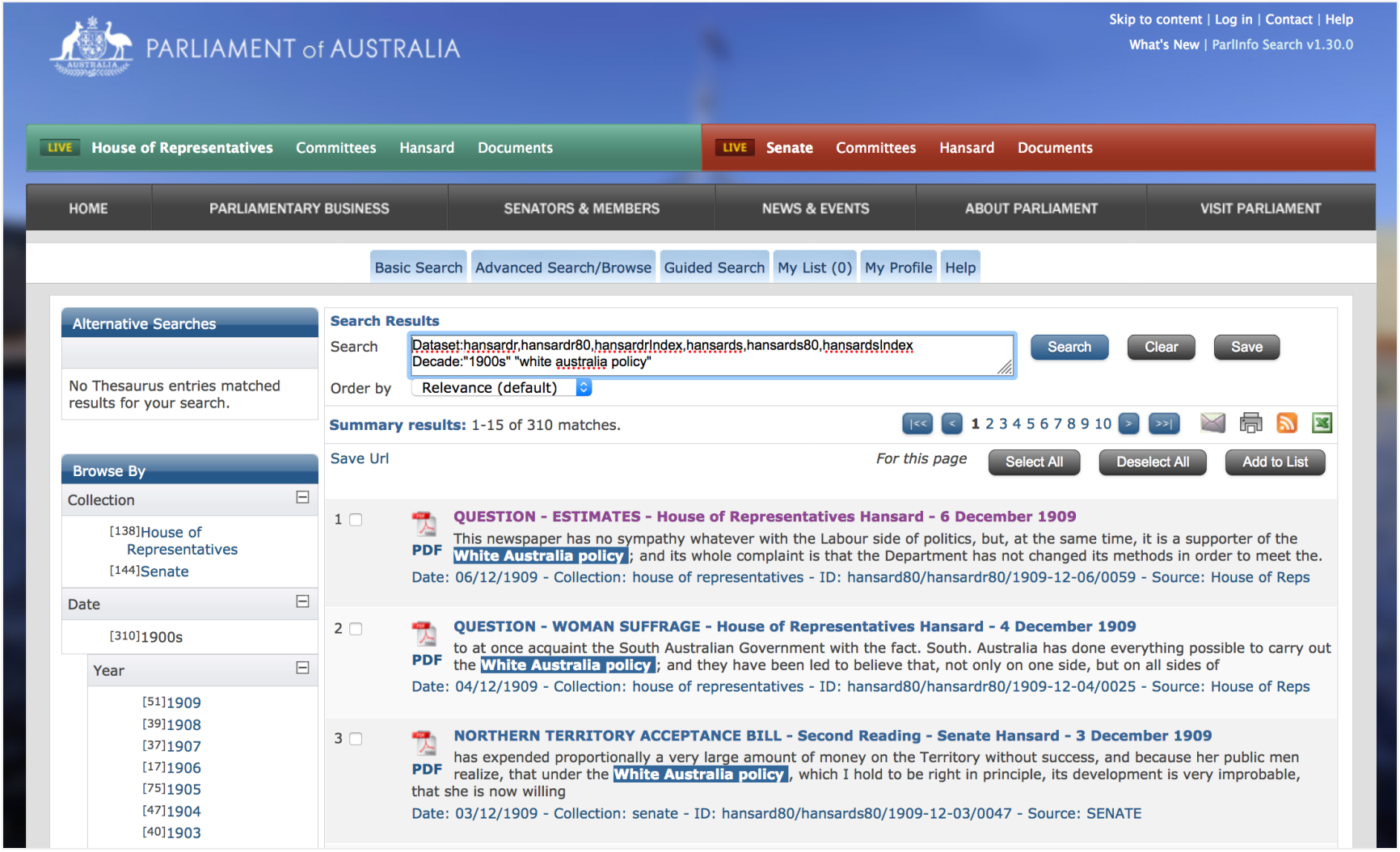


https://glam-workbench.github.io/csv-explorer/

is it in the form I need?

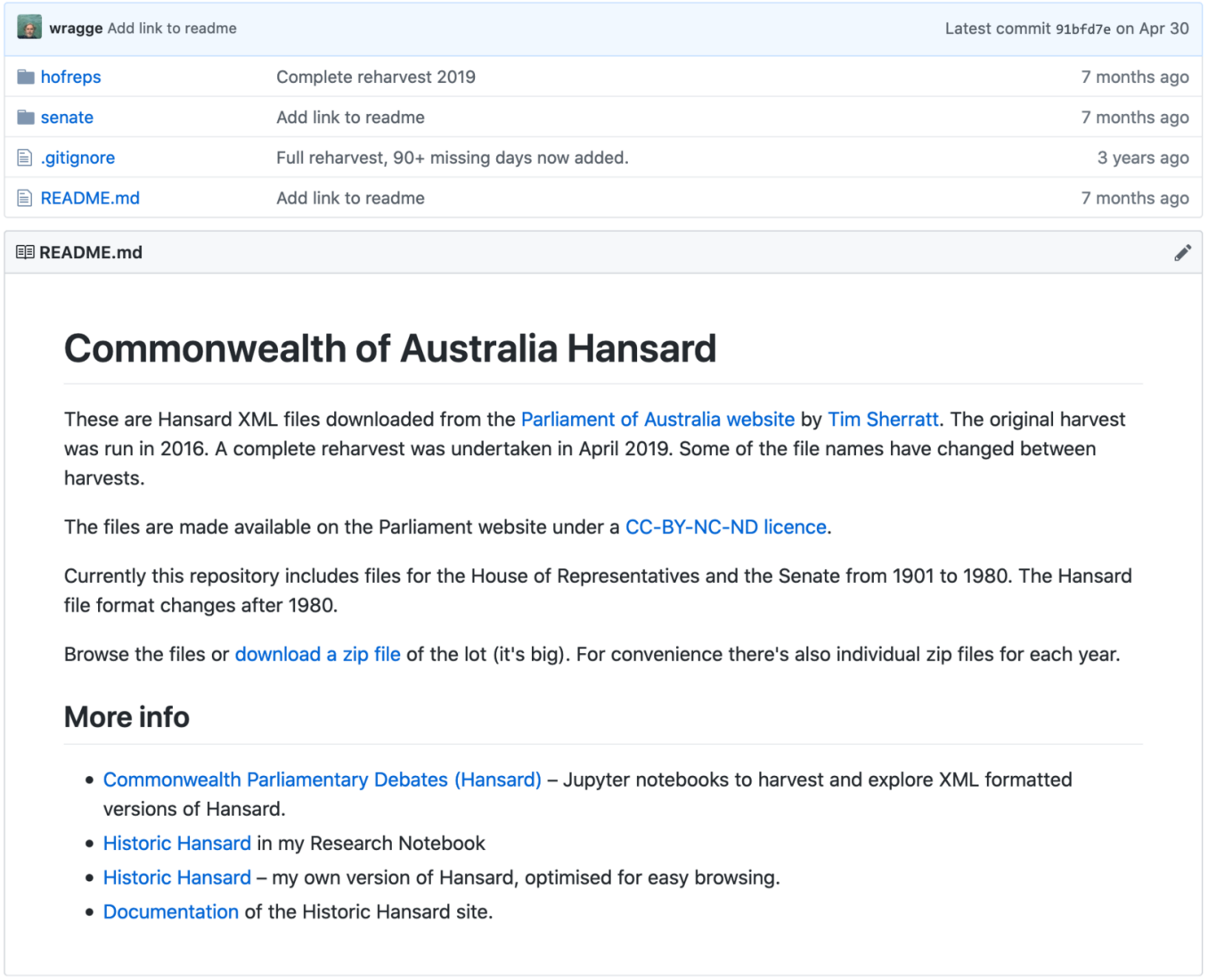
downloads are cool but…

spoken in Parliament



https://parlinfo.aph.gov.au/parlInfo/search/summary/summary.w3p

download XML in bulk



wragge Add link to readme — Latest commit 91bfd7e on Apr 30

hofreps	Complete reharvest 2019	7 months ago
senate	Add link to readme	7 months ago
.gitignore	Full reharvest, 90+ missing days now added.	3 years ago
README.md	Add link to readme	7 months ago

README.md

Commonwealth of Australia Hansard

These are Hansard XML files downloaded from the Parliament of Australia website by Tim Sherratt. The original harvest was run in 2016. A complete reharvest was undertaken in April 2019. Some of the file names have changed between harvests.

The files are made available on the Parliament website under a CC-BY-NC-ND licence.

Currently this repository includes files for the House of Representatives and the Senate from 1901 to 1980. The Hansard file format changes after 1980.

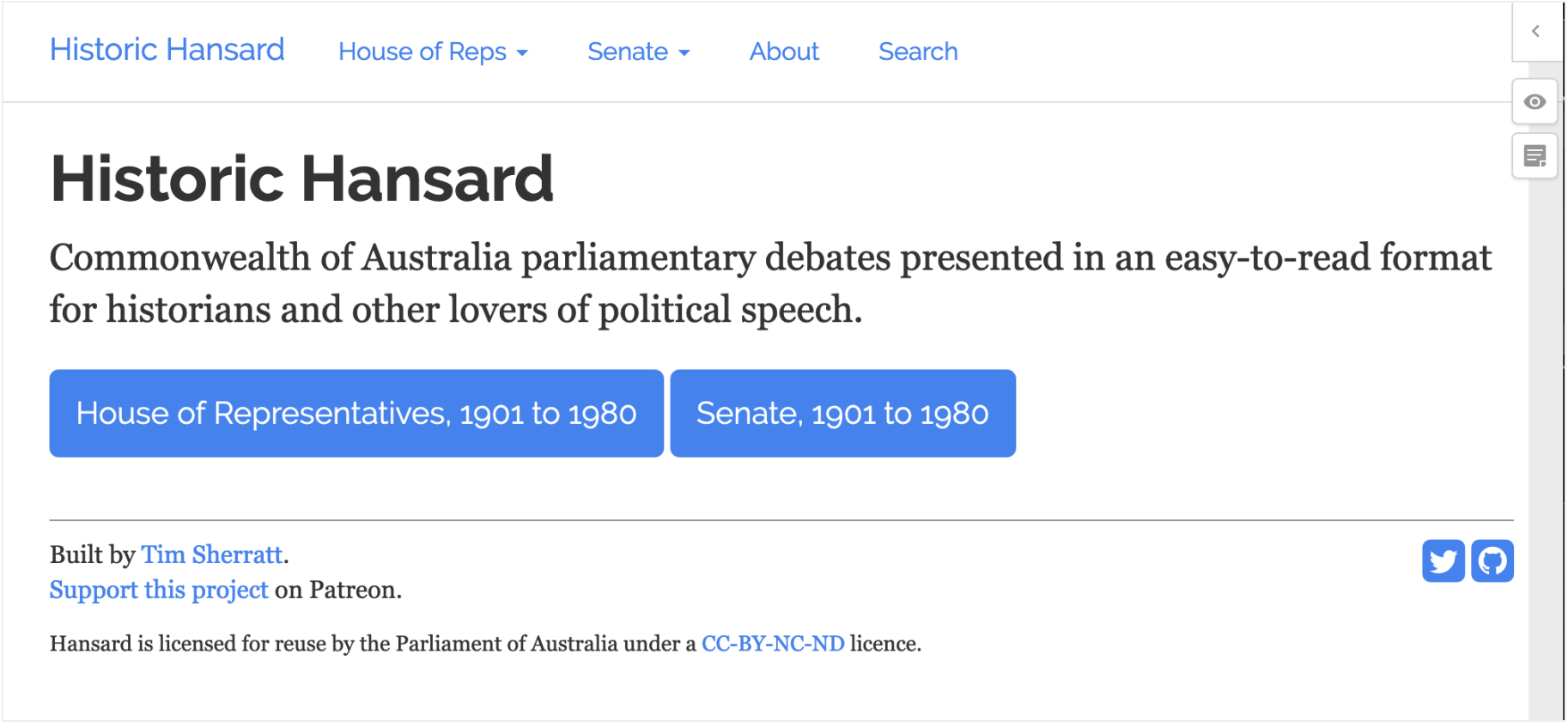
Browse the files or download a zip file of the lot (it's big). For convenience there's also individual zip files for each year.

More info

- Commonwealth Parliamentary Debates (Hansard) – Jupyter notebooks to harvest and explore XML formatted versions of Hansard.
- Historic Hansard in my Research Notebook
- Historic Hansard – my own version of Hansard, optimised for easy browsing.
- Documentation of the Historic Hansard site.

https://github.com/wragge/hansard-xml

one page per day



https://historichansard.net/

3,471 Bulletin editorial cartoons

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

> Previously this kind of research would have taken months but now thanks to Tim's shortcut can be done in a matter of days.
>
> — *Guy Hansen, National Library of Australia*

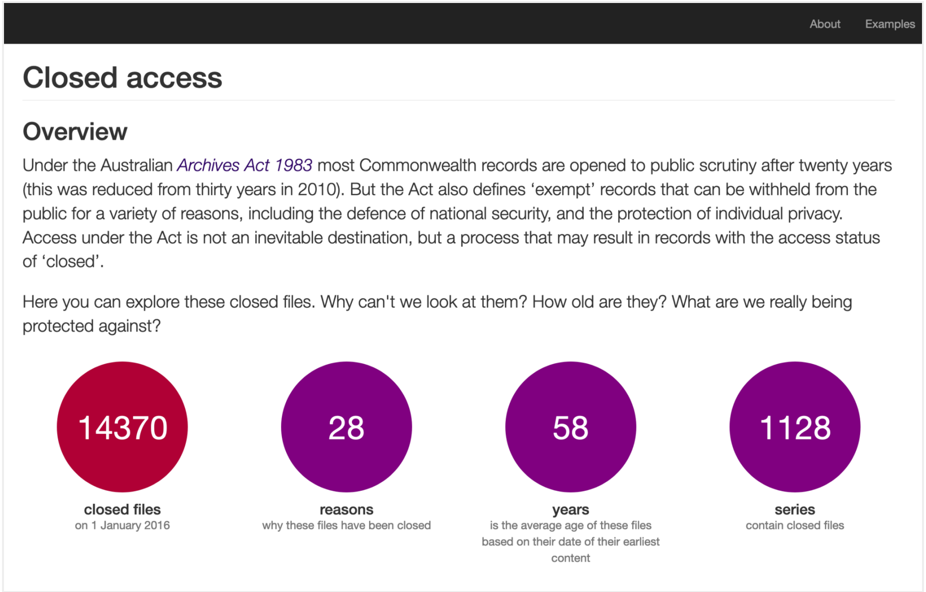
https://www.nla.gov.au/stories/audio/the-best-of-the-bulletin-cartoonists

scraping data from the NAA

series	total_items	date_from	date_to	Open	OWE	NYE	Closed	digitised_files	digitised_pages	% open	% digitised
B13	20,194	1800	2005	19,786	8	400	0	354	5,043	97.98%	1.75%
B6003	3	1904	1959	3	0	0	0	0	0	100.00%	0.00%
BP343/15	2,571	1916	1955	2,566	0	5	0	85	176	99.81%	3.31%
D2860	1	1902	1957	0	1	0	0	0	0	0.00%	0.00%
D5036	1	1906	1935	1	0	0	0	0	0	100.00%	0.00%
D596	11,395	1871	1971	2,983	31	8,381	0	185	3,031	26.18%	1.62%
E752	722	1905	1941	719	0	3	0	717	9,310	99.58%	99.31%
J2481	858	1897	1903	858	0	0	0	858	2,031	100.00%	100.00%
J2482	799	1902	1912	799	0	0	0	798	3,153	100.00%	99.87%
J2483	14,438	1903	1956	14,436	0	2	0	14,436	79,210	99.99%	99.99%
J3115	161	1899	1928	161	0	0	0	161	1,344	100.00%	100.00%
K1145	4,816	1900	1955	4,791	0	25	0	175	874	99.48%	3.63%
P437	4,958	1901	1940	4,945	10	2	1	18	442	99.74%	0.36%
P526	2	1909	1918	1	0	1	0	0	0	50.00%	0.00%
PP4/2	613	1903	1947	610	0	3	0	28	1,512	99.51%	4.57%
PP6/1	6,010	1906	1978	1,863	33	4,109	5	245	6,461	31.00%	4.08%
SP11/26	27	1902	1902	27	0	0	0	5	84	100.00%	18.52%
SP11/6	191	1902	1947	101	0	90	0	1	323	52.88%	0.52%
SP115/1	1,787	1884	1943	1,787	0	0	0	9	285	100.00%	0.50%
SP115/10	6	1884	1888	6	0	0	0	0	0	100.00%	0.00%
SP42/1	16,256	1881	1960	15,525	0	731	0	3,253	45,862	95.50%	20.01%
SP726/1	6	1902	1959	6	0	0	0	0	0	100.00%	0.00%
ST84/1	2,765	1855	1975	2,758	0	7	0	434	13,979	99.75%	15.70%

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


THE REAL FACE OF WHITE AUSTRALIA
MARK > TRANSCRIBE > VERIFY
ABOUT FIELD GUIDE BLOG DISCUSS
LOG IN TO SAVE YOUR WORK:
The Real Face of White Australia
THE REAL FACE OF WHITE AUSTRALIA
Join us in transcribing records that document the lives of ordinary people living under the restrictions of the White Australia Policy.
GET STARTED!

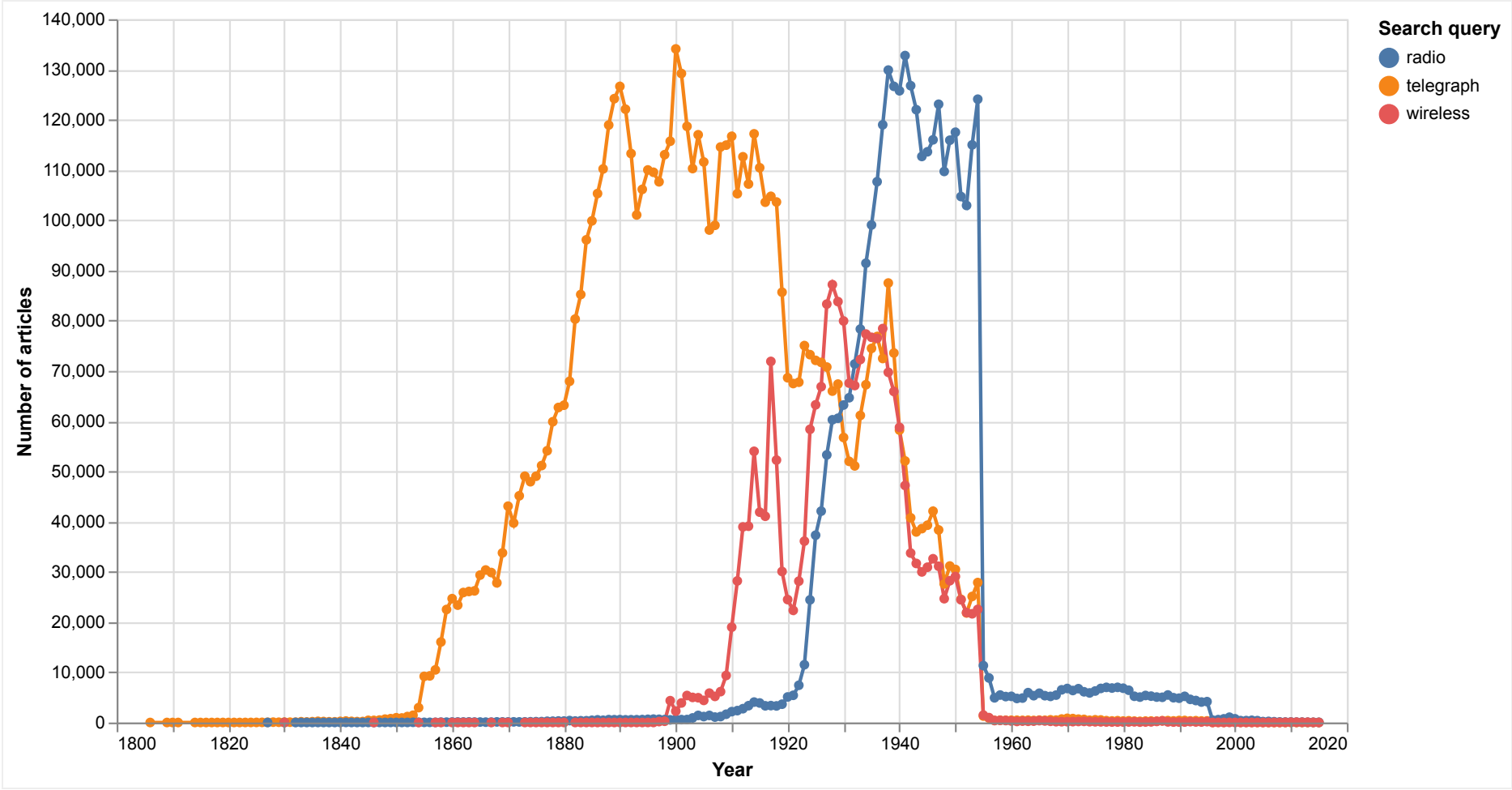

About Examples
Closed access
Overview
Under the Australian Archives Act 1983 most Commonwealth records are opened to public scrutiny after twenty years (this was reduced from thirty years in 2010). But the Act also defines 'exempt' records that can be withheld from the public for a variety of reasons, including the defence of national security, and the protection of individual privacy. Access under the Act is not an inevitable destination, but a process that may result in records with the access status of 'closed'.
Here you can explore these closed files. Why can't we look at them? How old are they? What are we really being protected against?
14370
closed files
on 1 January 2016
28
reasons
why these files have been closed
58
years
is the average age of these files based on their date of their earliest content
1128
series
contain closed files


redacted:
acts of erasure, censorship, and control extracted from [redacted] held by the
here voids become [redacted] on the practice of state surveillance
sampling 239571 redactions...

do I have the technical skills?

APIs are cool, but…

with the Trove API you can...



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

visualise newspaper searches

Visualise Trove newspaper searches over time

You know the feeling. You enter a query into Trove's digitised newspapers search box and...

Hmmm, **3 million results**, how do you make sense of that..?

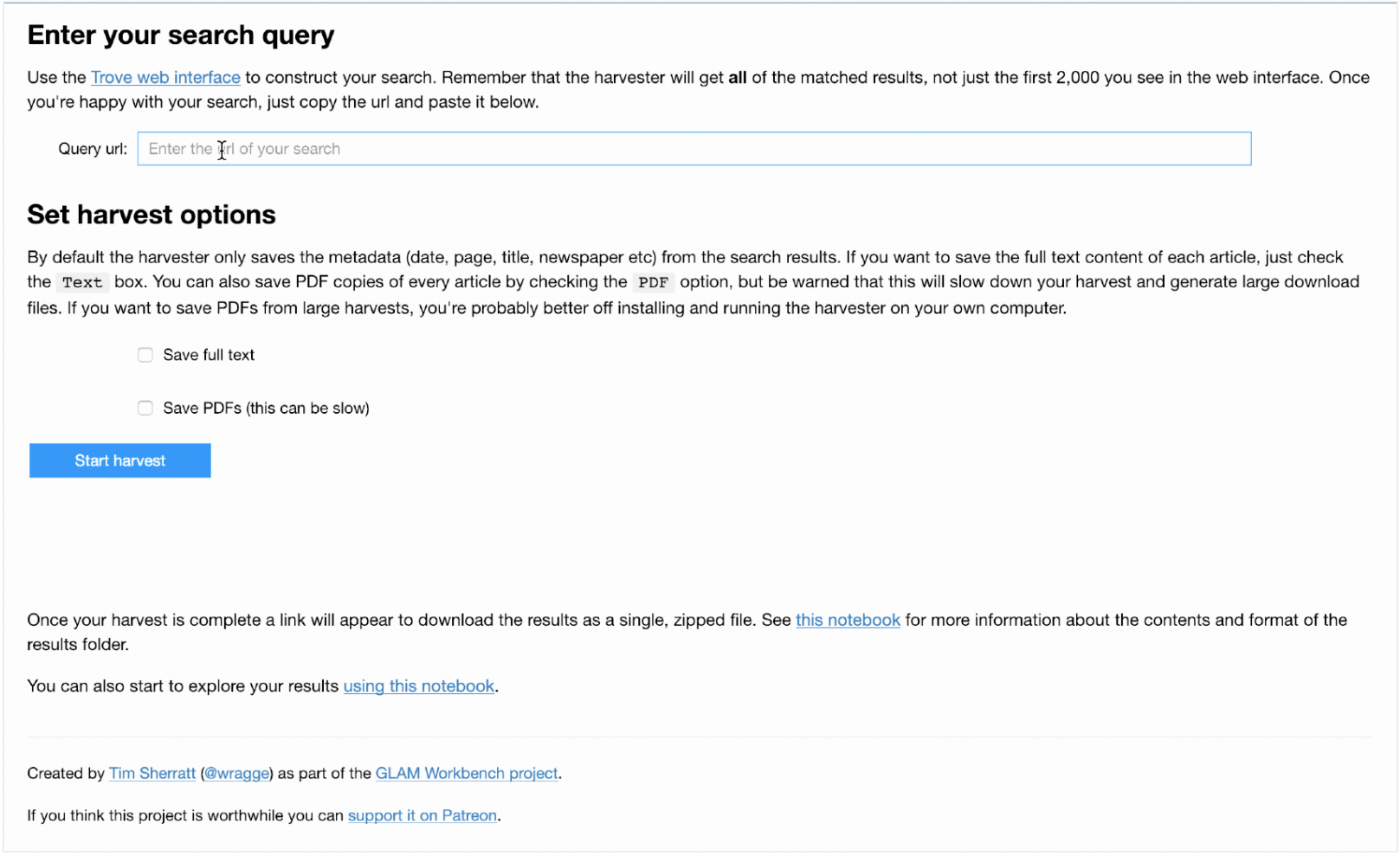
Trove tries to be as helpful as possible by ordering your results by relevance. This is great if you aim is to find a few interesting articles. But how can you get a sense of the complete results set? How can you *see* everything? Trove's web interface only shows you the first 2,000 articles matching your search. But by getting data directly from the Trove API we can go bigger.

This notebook helps you zoom out and explore how the number of newspaper articles in your results varies over time by using the `decade` and `year` facets. We'll then combine this approach with other search facets to see how we can slice a set of results up in different ways to investigate historical changes.

If you're interested in exploring the possibilities examined in this notebook, but are feeling a bit intimidated by the code, skip to the Related resources section for some alternative starting points. But once you've got a bit of confidence, please come back here to learn more about how it all works!

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

get your newspaper articles in bulk



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

collected tools, examples & hacks

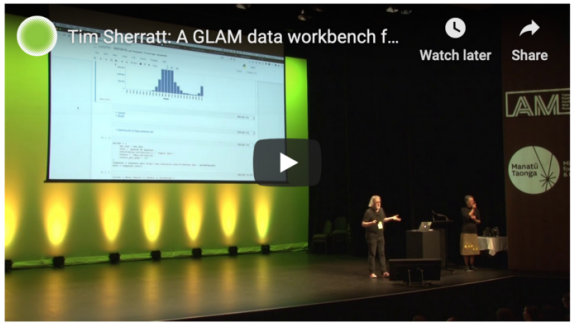
GLAM Workbench | Search | GLAM-Workbench 29 Repositories

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.

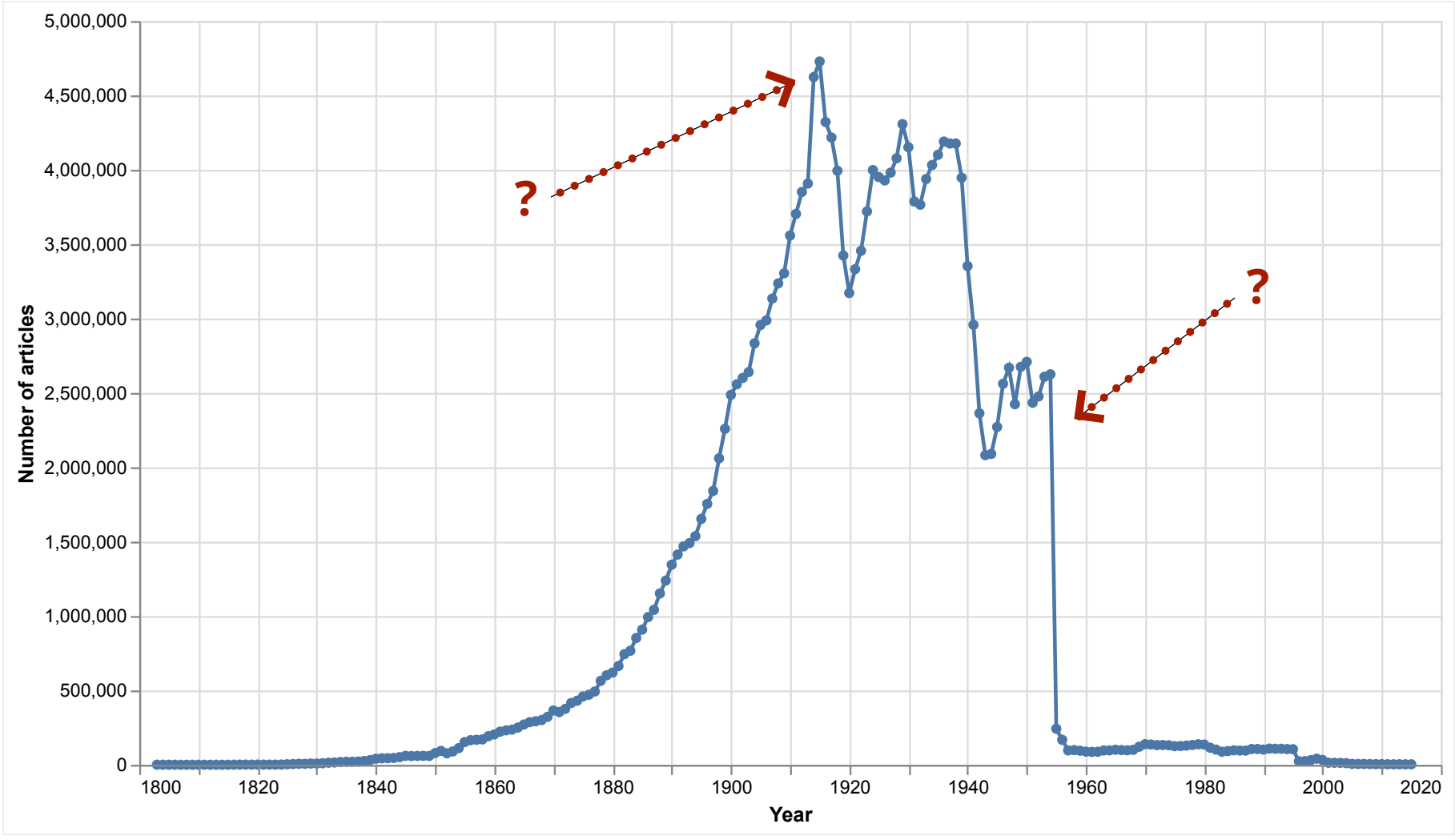
Table of contents

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

what does it really mean?

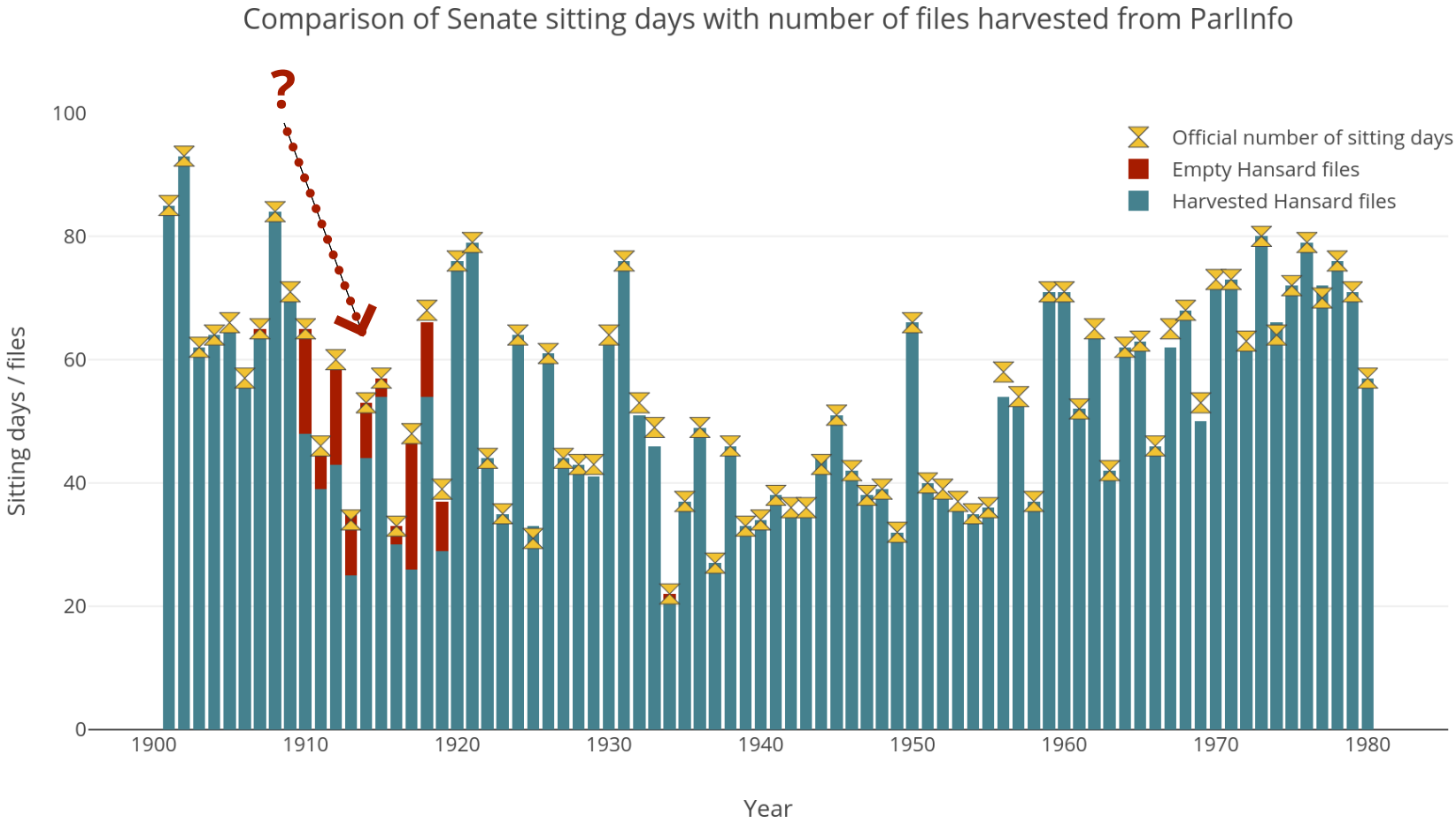
seeing differently....

the WWI effect



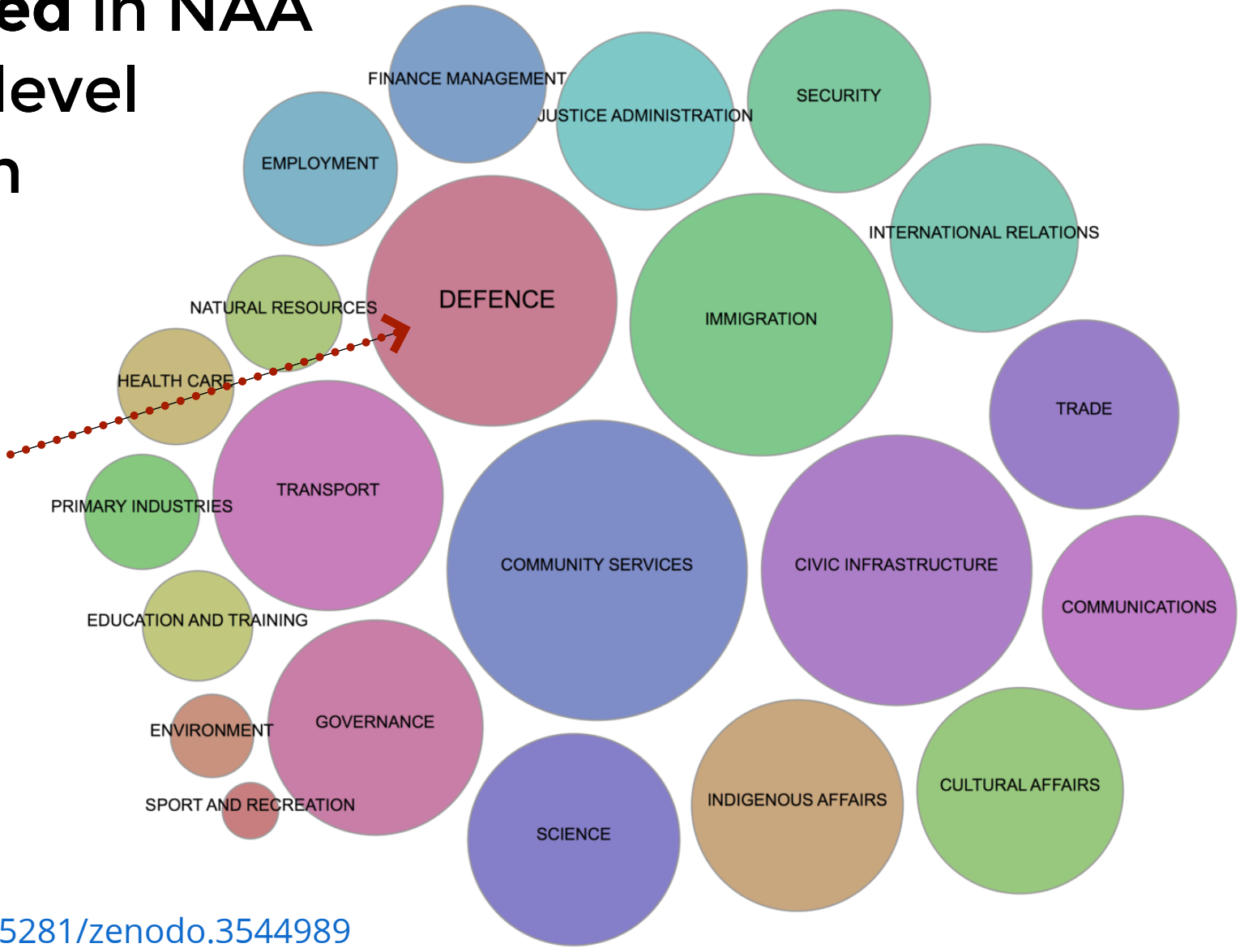
https://glam-workbench.github.io/trove-newspapers/

the Hansard black hole



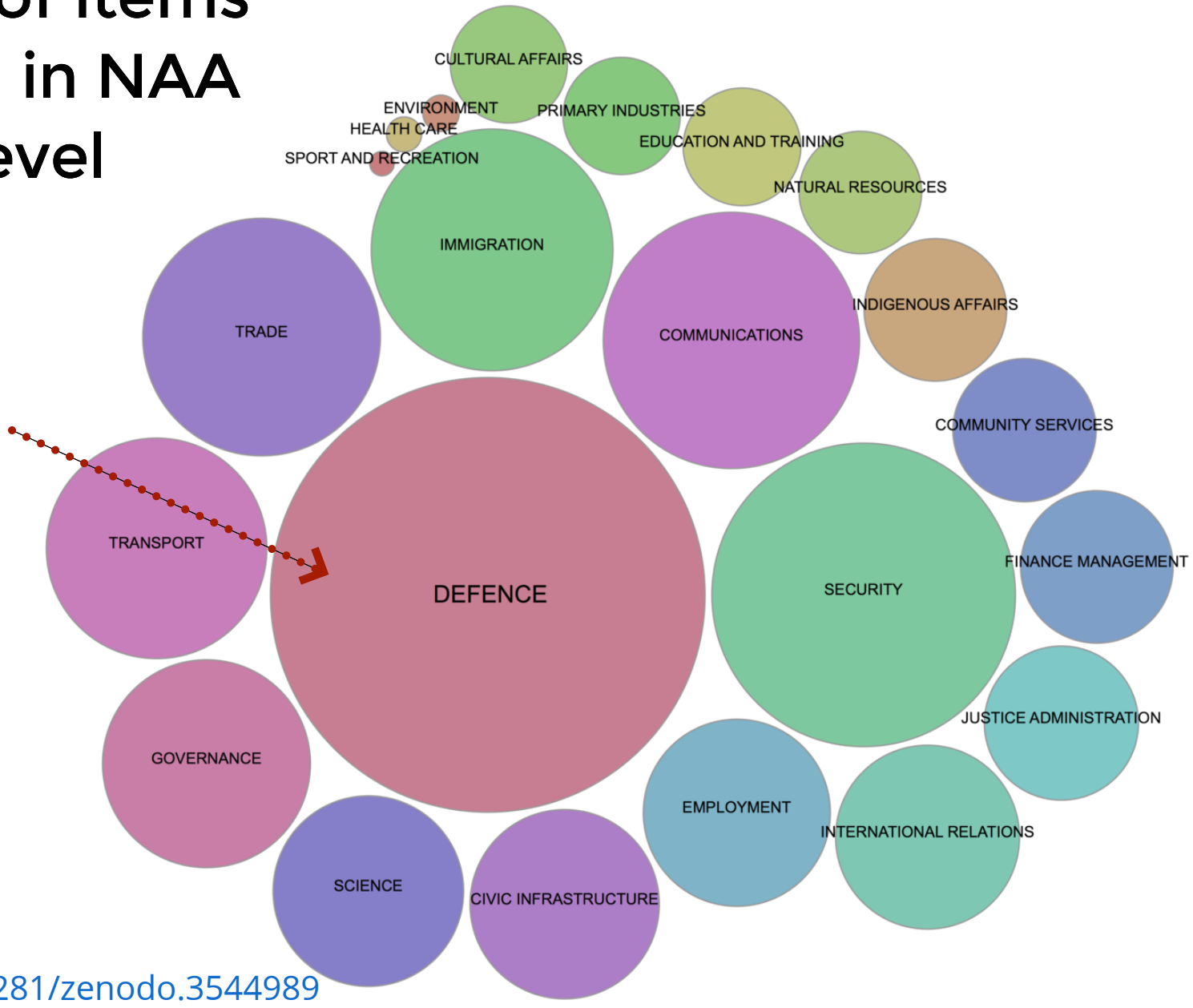
http://timsherratt.org/research-notebook/historic-hansard/notes/investigating-the-hansard-black-hole/

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



https://doi.org/10.5281/zenodo.3544989

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



https://doi.org/10.5281/zenodo.3544989

access is constructed

- how does access change?
- what can I see differently?
- what is possible that wasn't before?
- where can I go next?

HOME BLOG ABOUT UPDATES PROJECTS

Tim Sherratt

Historian and hacker

I'm a historian and hacker who researches the possibilities and politics of digital cultural collections. You can follow what I'm up though my updates or on my blog.

One day a week I'm Associate Professor of Digital Heritage at the University of Canberra. The rest of the time I'm available to work on your projects – so feel free to get in touch!

My main project at the moment is the GLAM Workbench, which brings together many examples, tools, code and tutorials to help people explore the digital collections of galleries, libraries, archives, and museums.

You can find some of my other work in my open research notebook and digital heritage handbook.

If you like what I do you can support me on Patreon.

https://timsherratt.org/