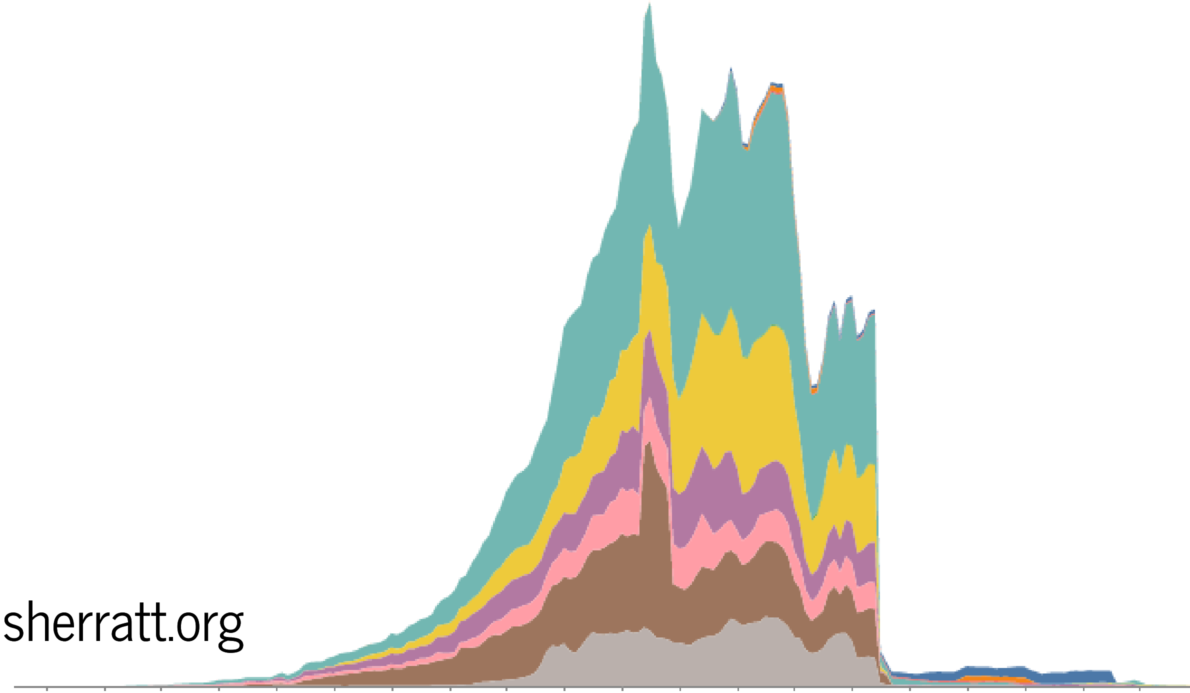
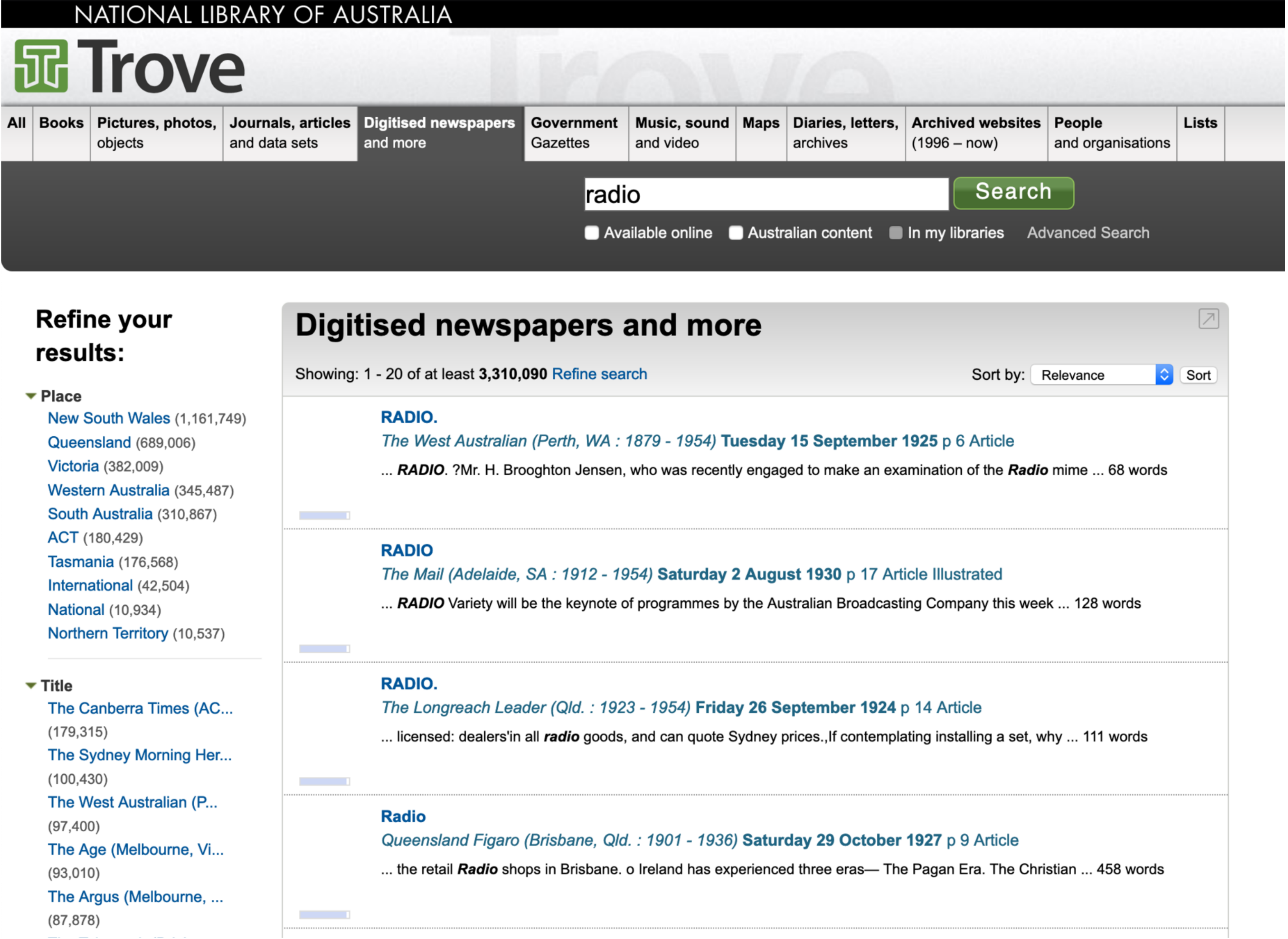
GLAM Collections as Data



Tim Sherratt · @wragge · timsherratt.org

These slides: https://slides.com/wragge/hussdata/

Watch the video: https://vimeo.com/350981680



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

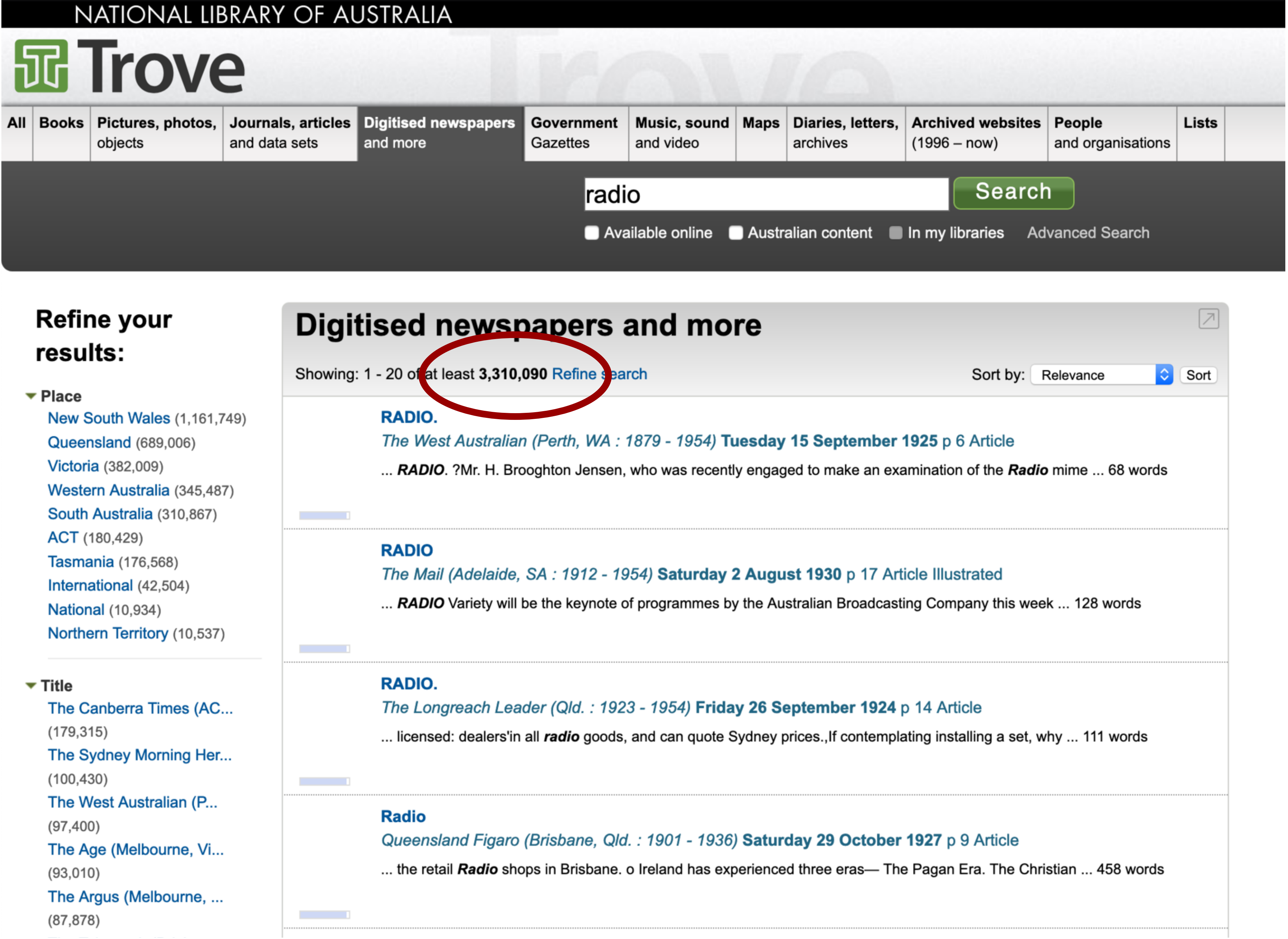
Showing: 1 - 20 of at least **3,310,090** Refine search

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

RADIO.
The West Australian (Perth, WA : 1879 - 1954) **Tuesday 15 September 1925** p 6 Article
... ***RADIO***. ?Mr. H. Broughton Jensen, who was recently engaged to make an examination of the ***Radio*** mime ... 68 words

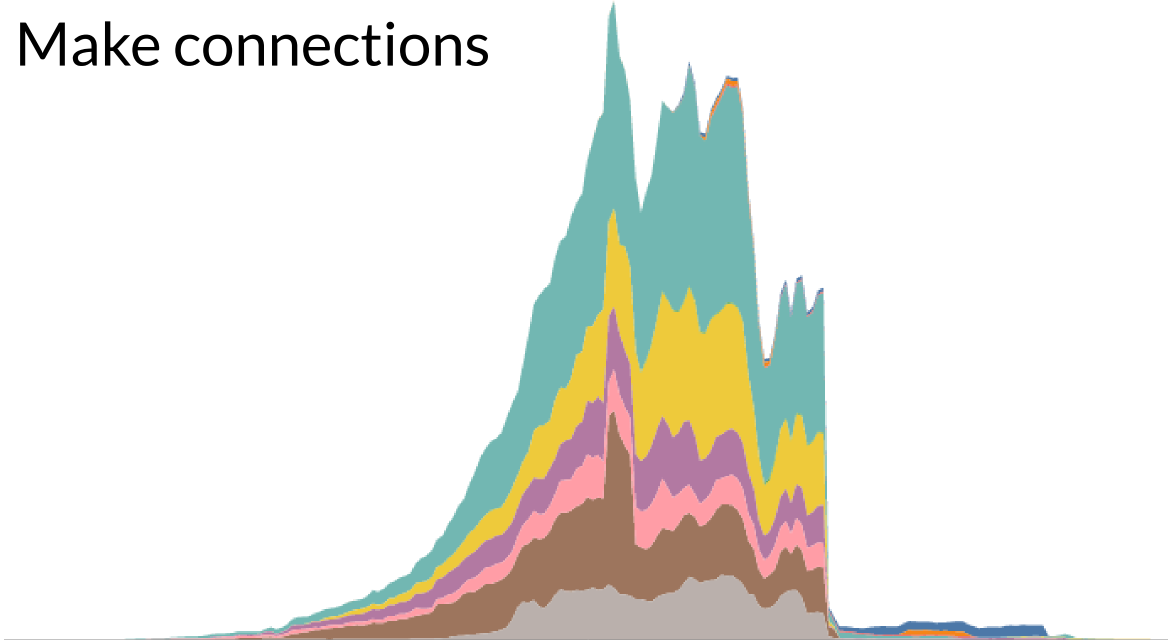
RADIO
The Mail (Adelaide, SA : 1912 - 1954) **Saturday 2 August 1930** p 17 Article Illustrated
... ***RADIO*** Variety will be the keynote of programmes by the Australian Broadcasting Company this week ... 128 words

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

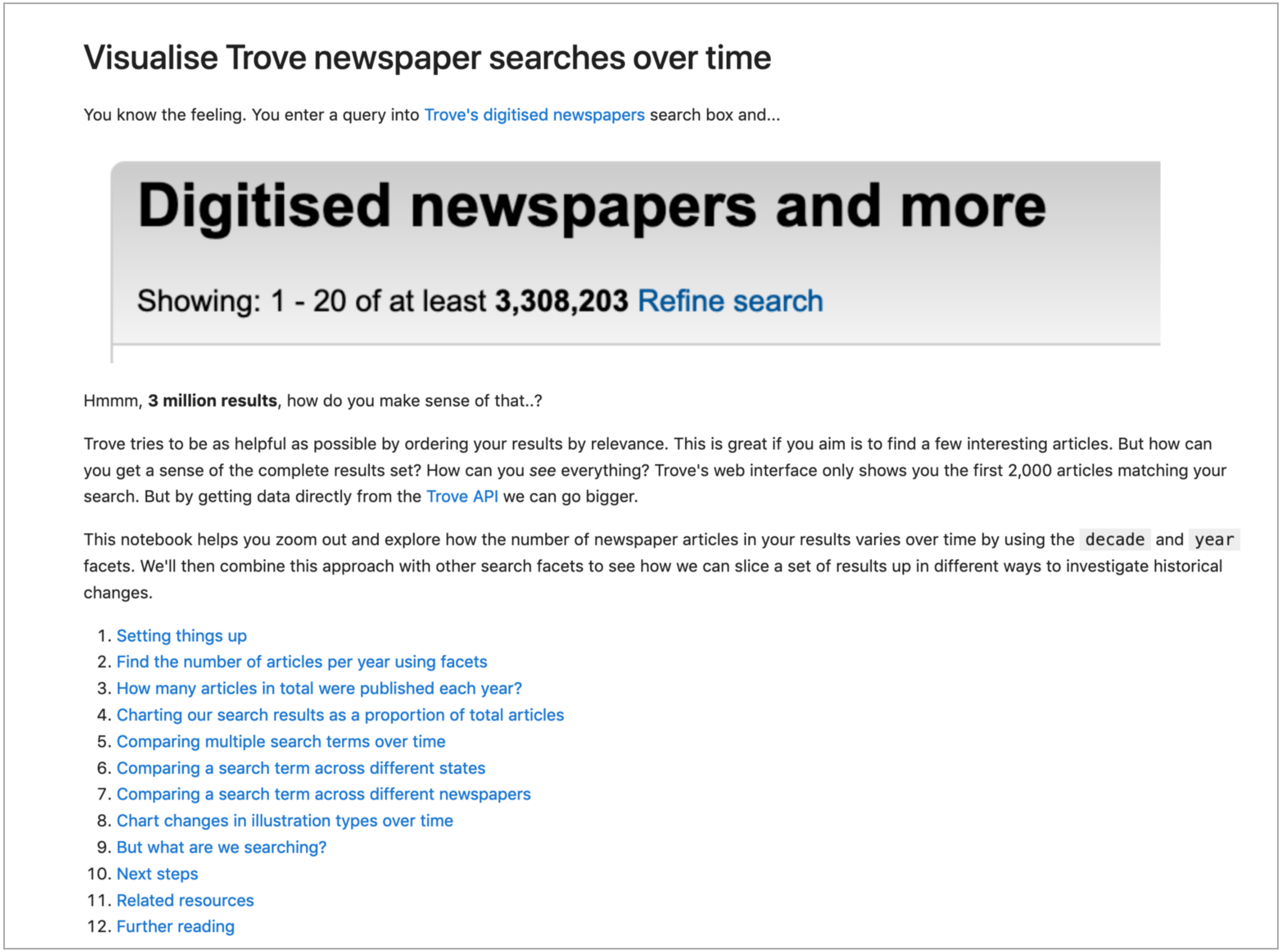
Collections as data

- Shift scales
- Find patterns
- Extract features
- Make connections



https://collectionsasdata.github.io/

https://collectionsasdata.github.io/



Visualise Trove newspaper searches over time

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

Digitised newspapers and more

Showing: 1 - 20 of at least **3,308,203** Refine search

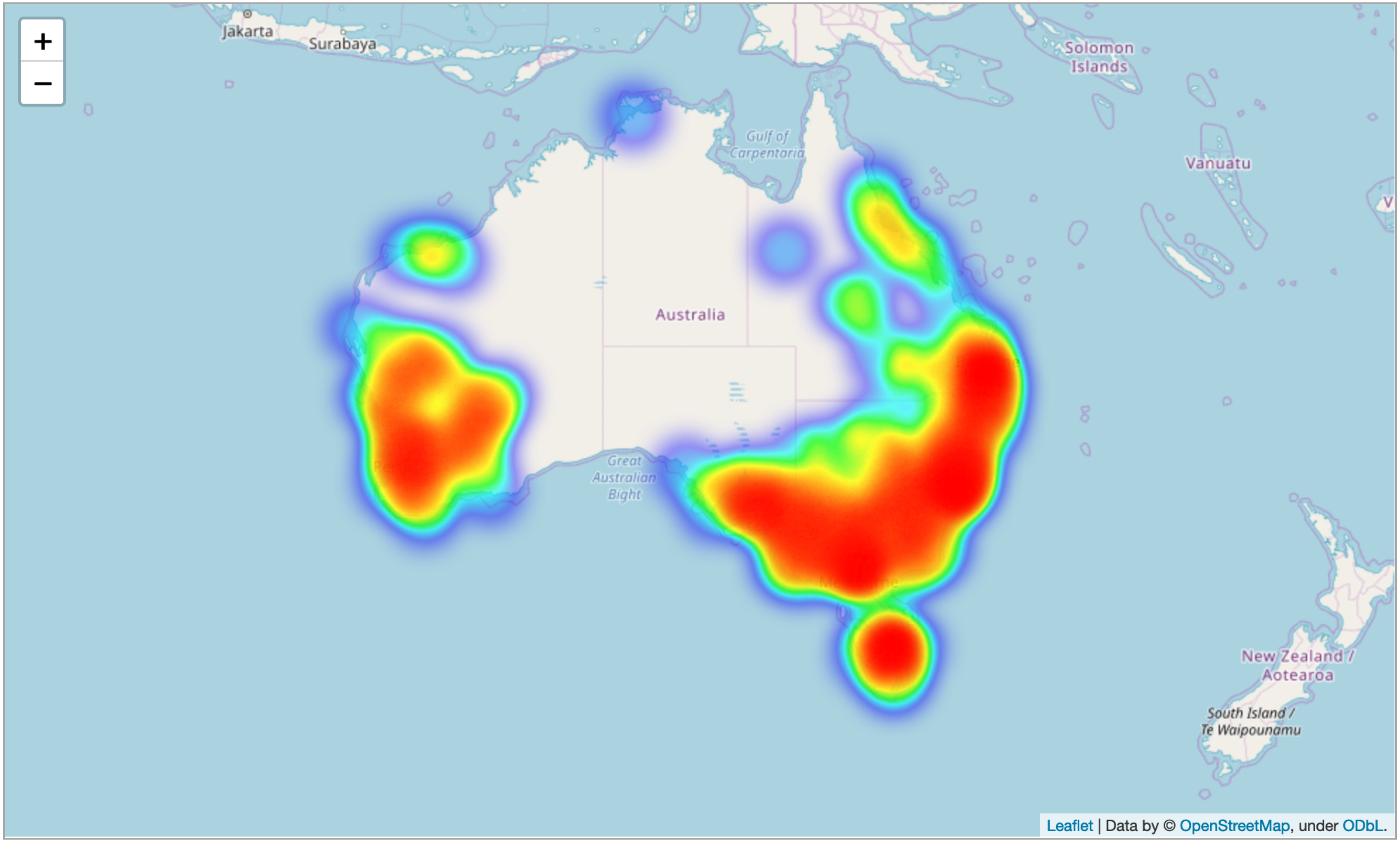
Hmmm, **3 million results**, how do you make sense of that..?

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.

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

Trove newspaper search by place of publication



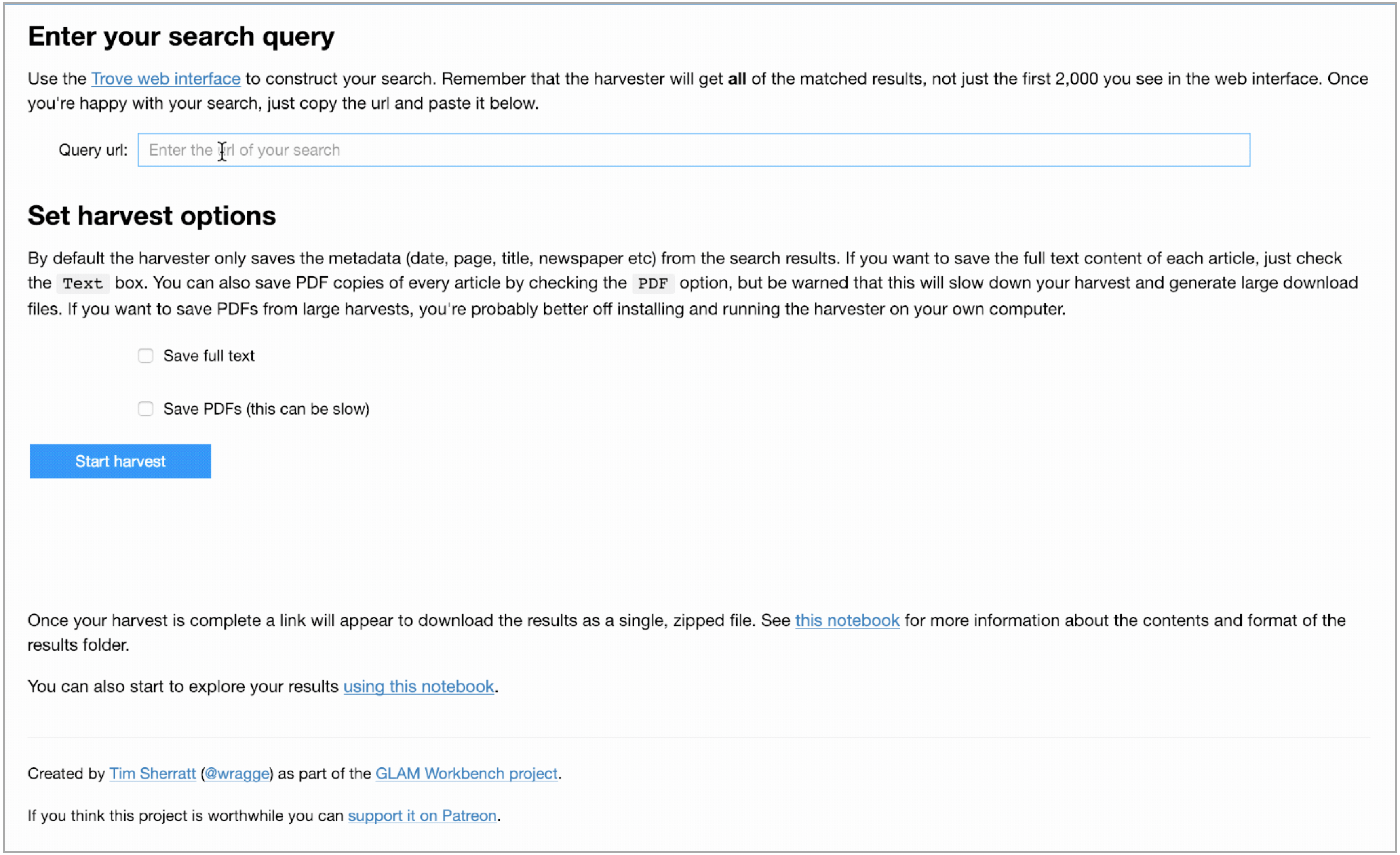
https://glam-workbench.github.io/trove-newspapers/

Newspaper articles with 'White Australia Policy' in their title

https://easyzoom.com/embed/139535

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

Harvest newspaper articles in bulk



Enter your search query

Use the Trove web interface to construct your search. Remember that the harvester will get **all** of the matched results, not just the first 2,000 you see in the web interface. Once you're happy with your search, just copy the url and paste it below.

Query url: Enter the url of your search

Set harvest options

By default the harvester only saves the metadata (date, page, title, newspaper etc) from the search results. If you want to save the full text content of each article, just check the `Text` box. You can also save PDF copies of every article by checking the `PDF` option, but be warned that this will slow down your harvest and generate large download files. If you want to save PDFs from large harvests, you're probably better off installing and running the harvester on your own computer.

□ Save full text

□ Save PDFs (this can be slow)

Once your harvest is complete a link will appear to download the results as a single, zipped file. See this notebook for more information about the contents and format of the results folder.

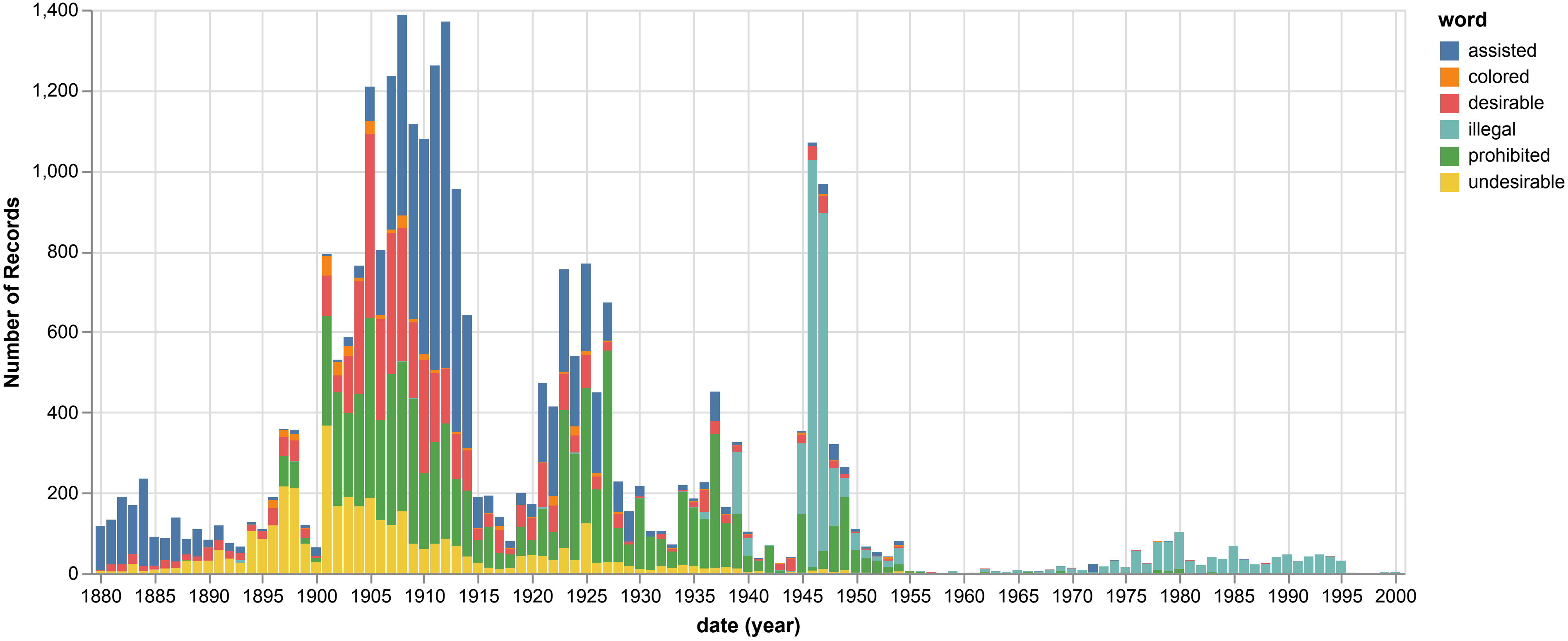
You can also start to explore your results using this notebook.

Created by Tim Sherratt (@wragge) as part of the GLAM Workbench project.

If you think this project is worthwhile you can support it on Patreon.

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

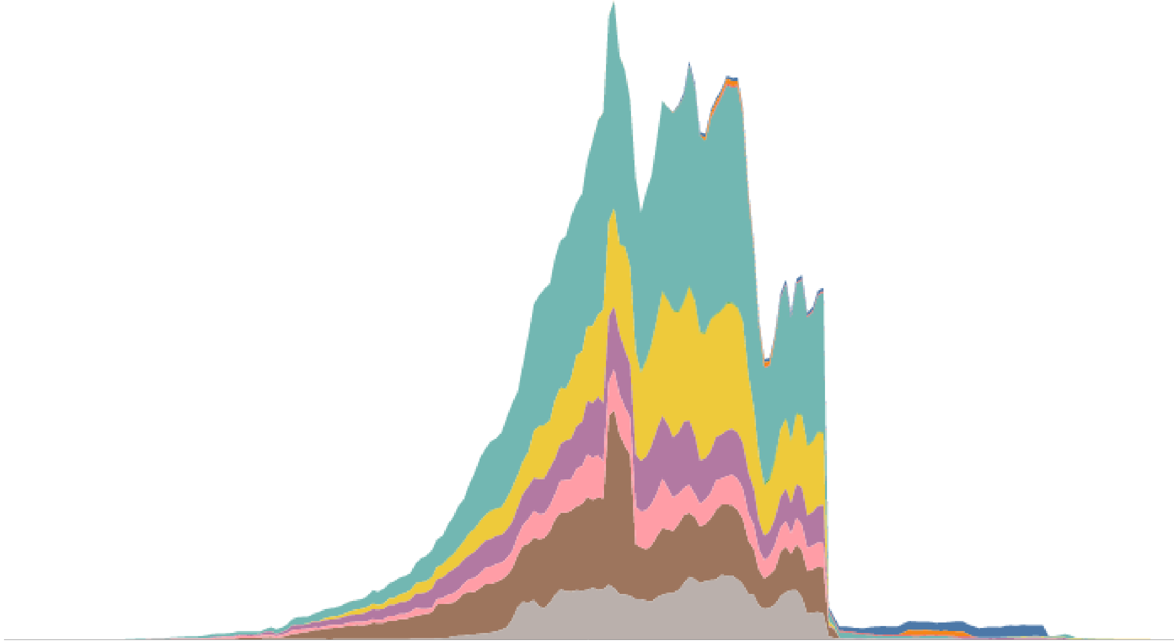
Words before 'aliens' in Australian newspapers



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

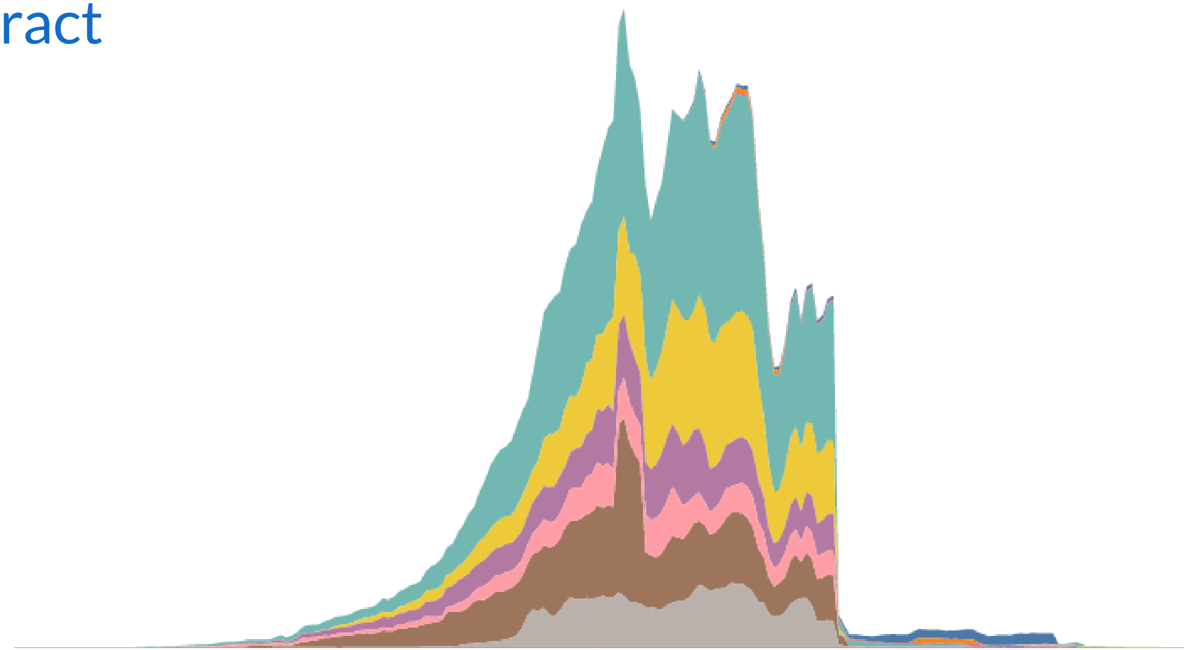
Jupyter notebooks

- computational narratives
- experimentation & learning
- tools & tutorials

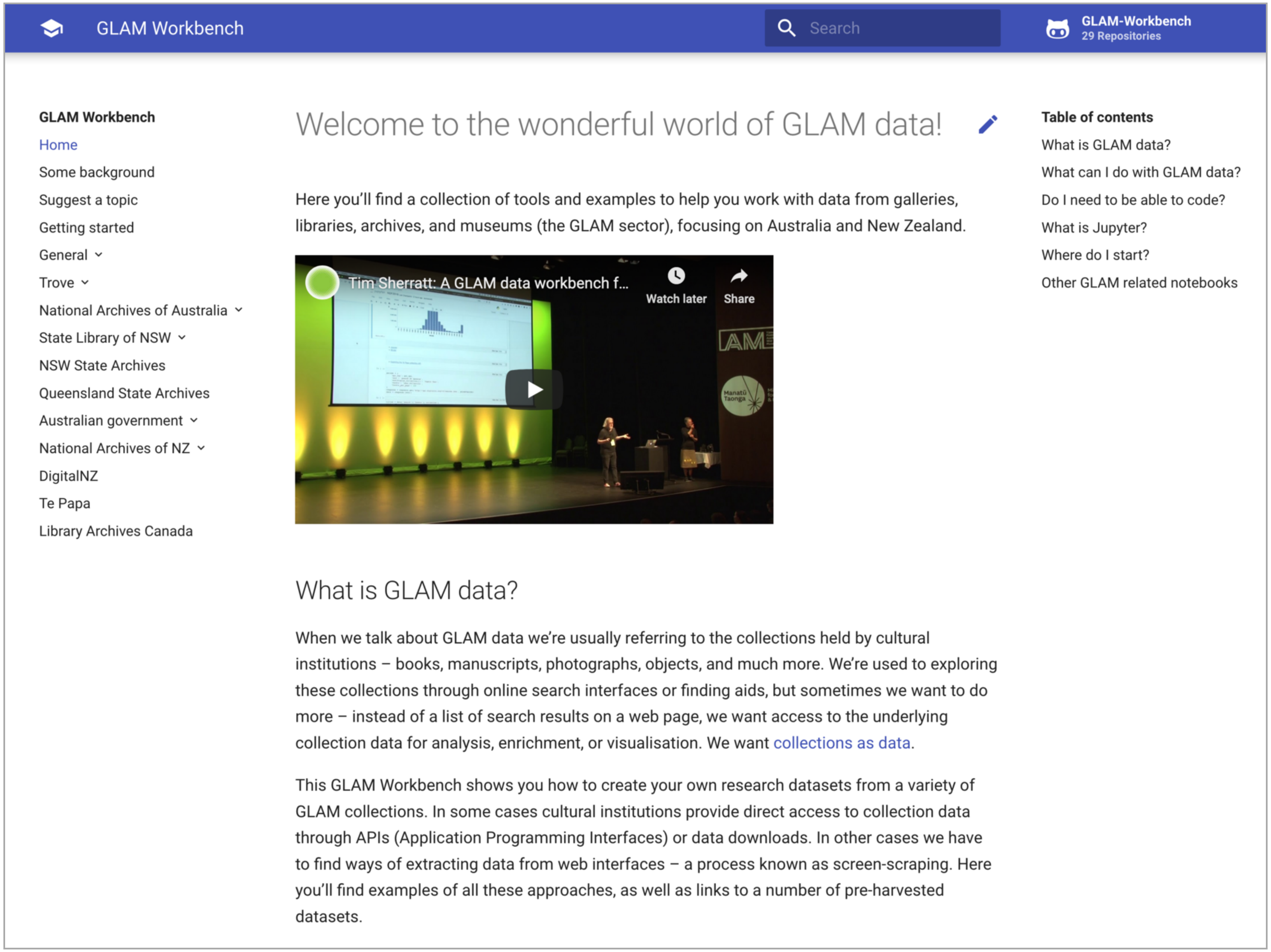


In the cloud or on your desktop

- Binder
- Tinker
- SWAN (part of Cloudstor)
- CoLab
- nteract



GLAM Workbench



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.

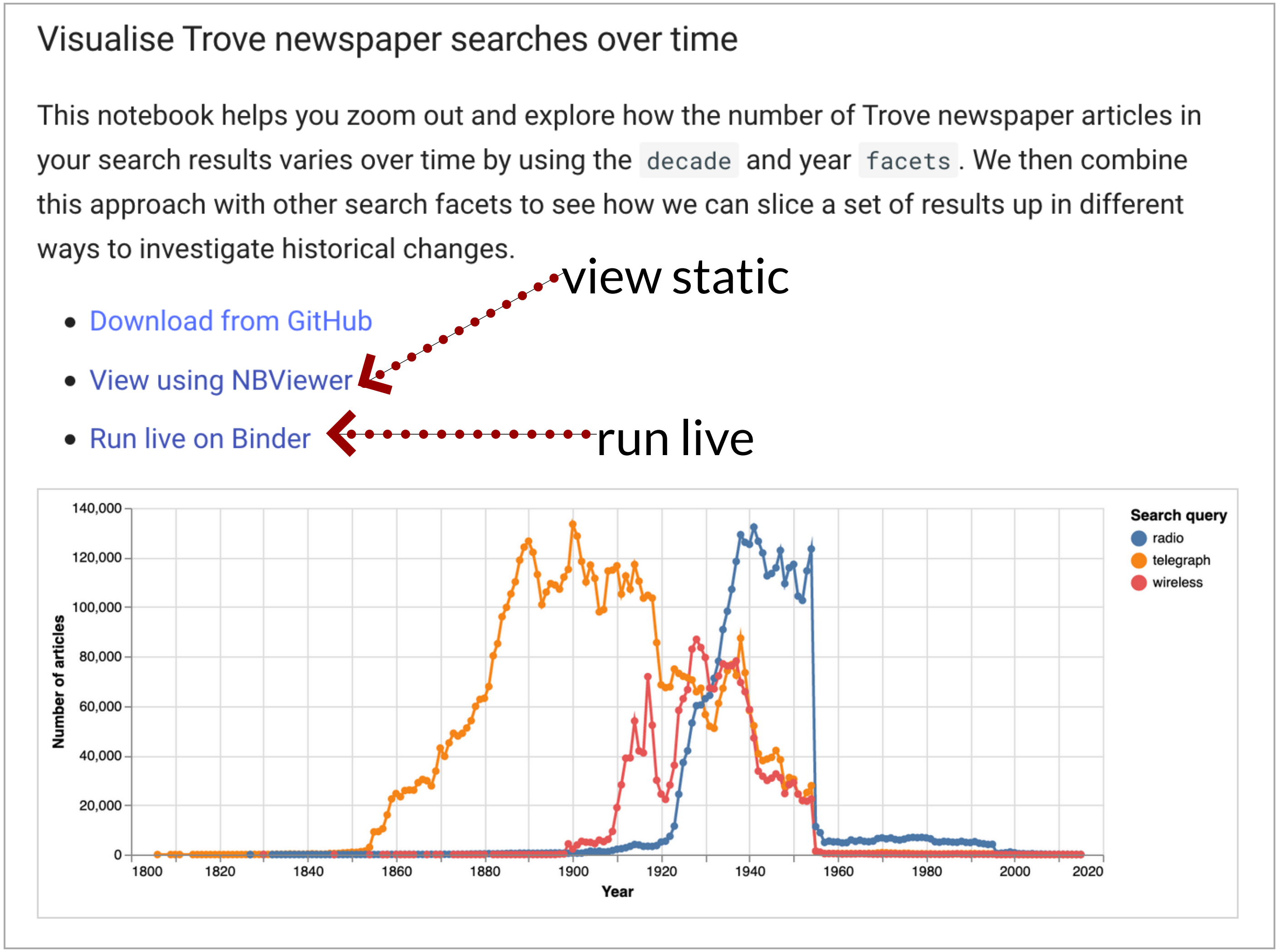
What is GLAM data?

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.

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

Static or live



Visualise Trove newspaper searches over time

This notebook helps you zoom out and explore how the number of Trove newspaper articles in your search results varies over time by using the `decade` and year `facets`. We 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.

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

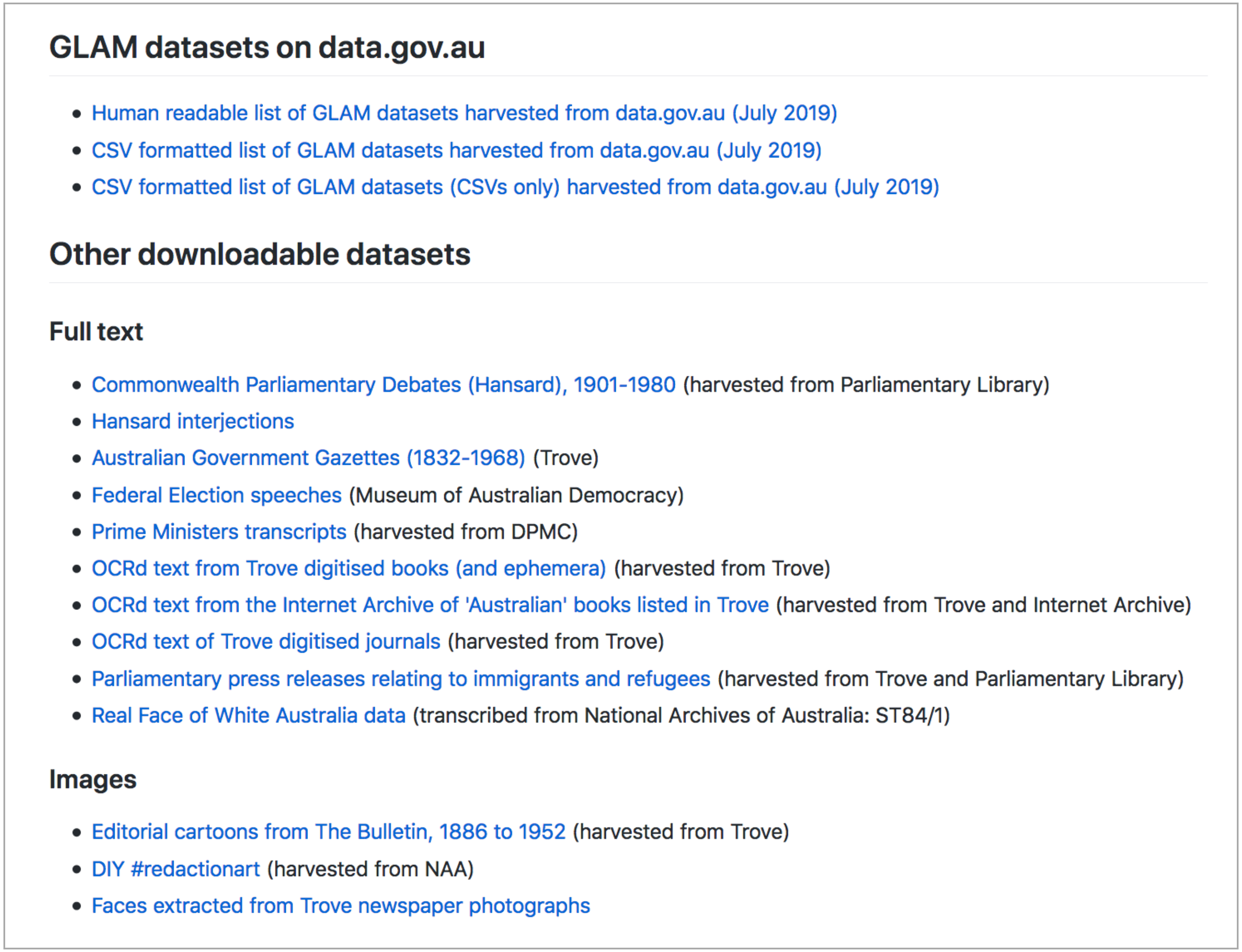
view static

run live

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

What data?

Sources of GLAM data



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)

Images

- Editorial cartoons from The Bulletin, 1886 to 1952 (harvested from Trove)
- DIY #redactionart (harvested from NAA)
- Faces extracted from Trove newspaper photographs

https://github.com/GLAM-Workbench/glam-data-list

- Machine-readable but coding required for download
- Downloadable but not easily findable
- Structured but not downloadable
- Downloadable but in many separate pieces

Machine-readable but coding required for download

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

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

Machine-readable but coding required for download

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

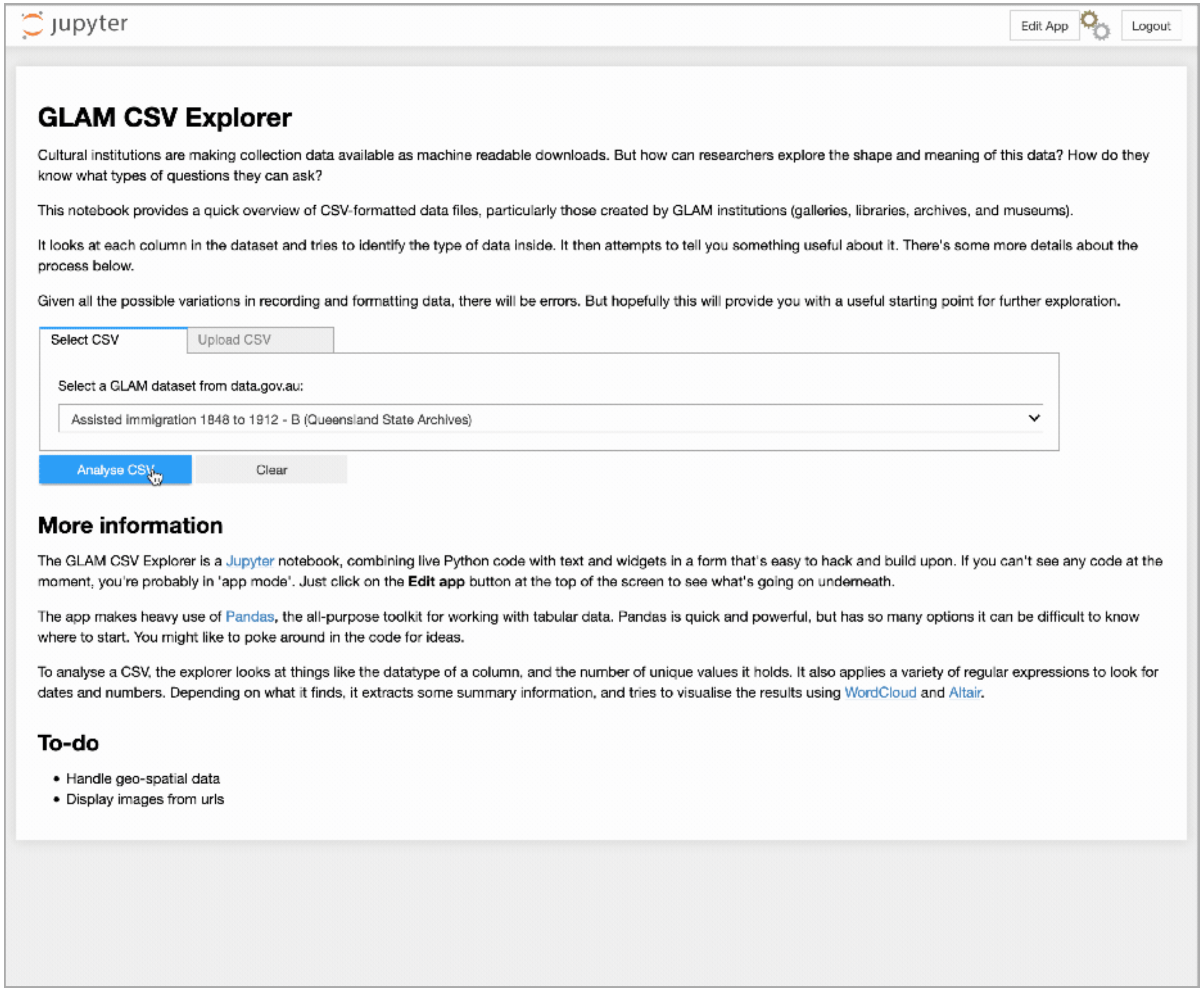
https://glam-workbench.github.io/digitalnz/

Downloadable but not easily findable

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

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

Downloadable but not easily findable



GLAM CSV Explorer

Cultural institutions are making collection data available as machine readable downloads. But how can researchers explore the shape and meaning of this data? How do they know what types of questions they can ask?

This notebook provides a quick overview of CSV-formatted data files, particularly those created by GLAM institutions (galleries, libraries, archives, and museums).

It looks at each column in the dataset and tries to identify the type of data inside. It then attempts to tell you something useful about it. There's some more details about the process below.

Given all the possible variations in recording and formatting data, there will be errors. But hopefully this will provide you with a useful starting point for further exploration.

Select a GLAM dataset from data.gov.au:

Assisted Immigration 1848 to 1912 - B (Queensland State Archives)

More information

The GLAM CSV Explorer is a Jupyter notebook, combining live Python code with text and widgets in a form that's easy to hack and build upon. If you can't see any code at the moment, you're probably in 'app mode'. Just click on the **Edit app** button at the top of the screen to see what's going on underneath.

The app makes heavy use of Pandas, the all-purpose toolkit for working with tabular data. Pandas is quick and powerful, but has so many options it can be difficult to know where to start. You might like to poke around in the code for ideas.

To analyse a CSV, the explorer looks at things like the datatype of a column, and the number of unique values it holds. It also applies a variety of regular expressions to look for dates and numbers. Depending on what it finds, it extracts some summary information, and tries to visualise the results using WordCloud and Altair.

To-do

- Handle geo-spatial data
- Display images from urls

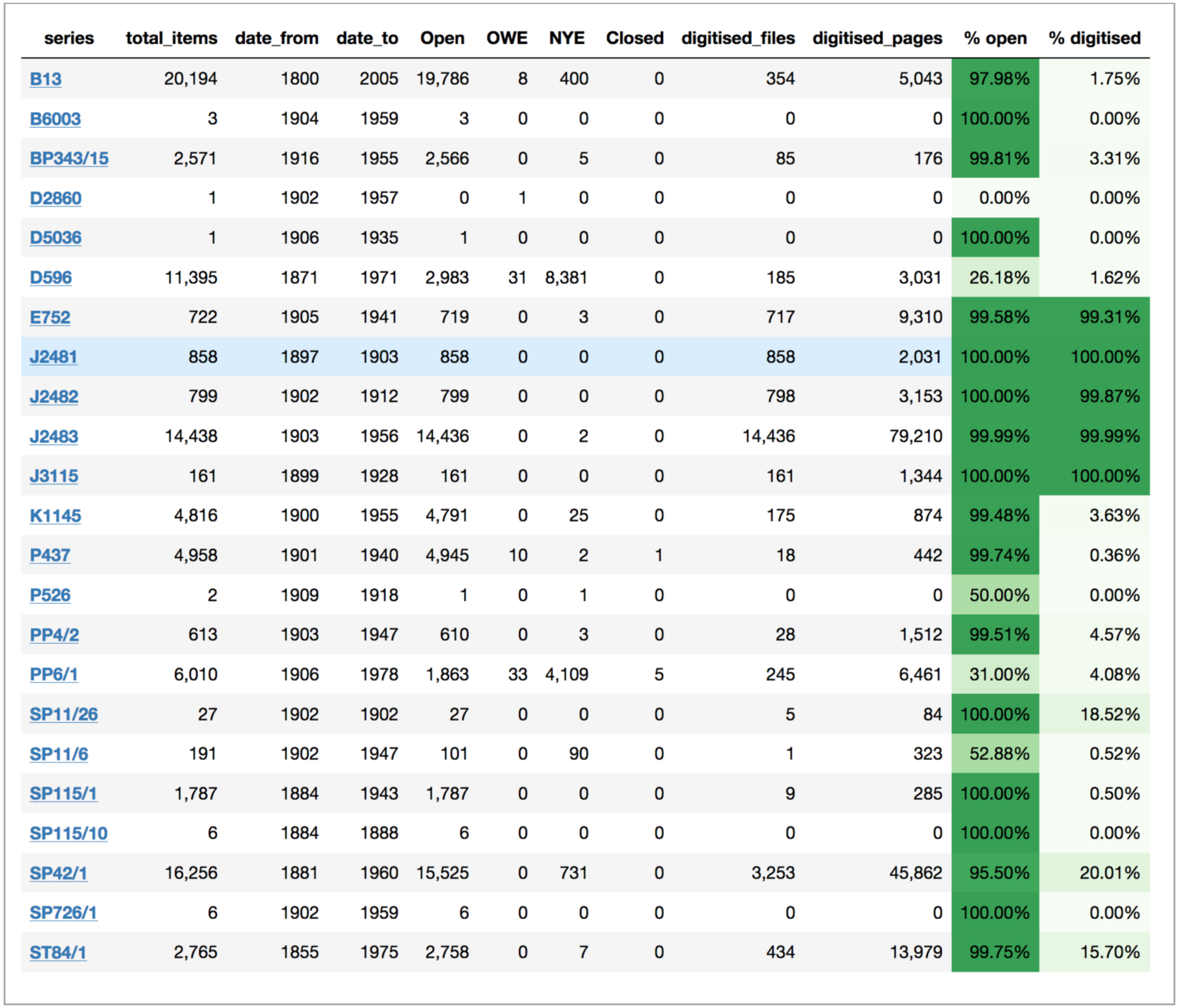
https://glam-workbench.github.io/csv-explorer/

Structured but not downloadable

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

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

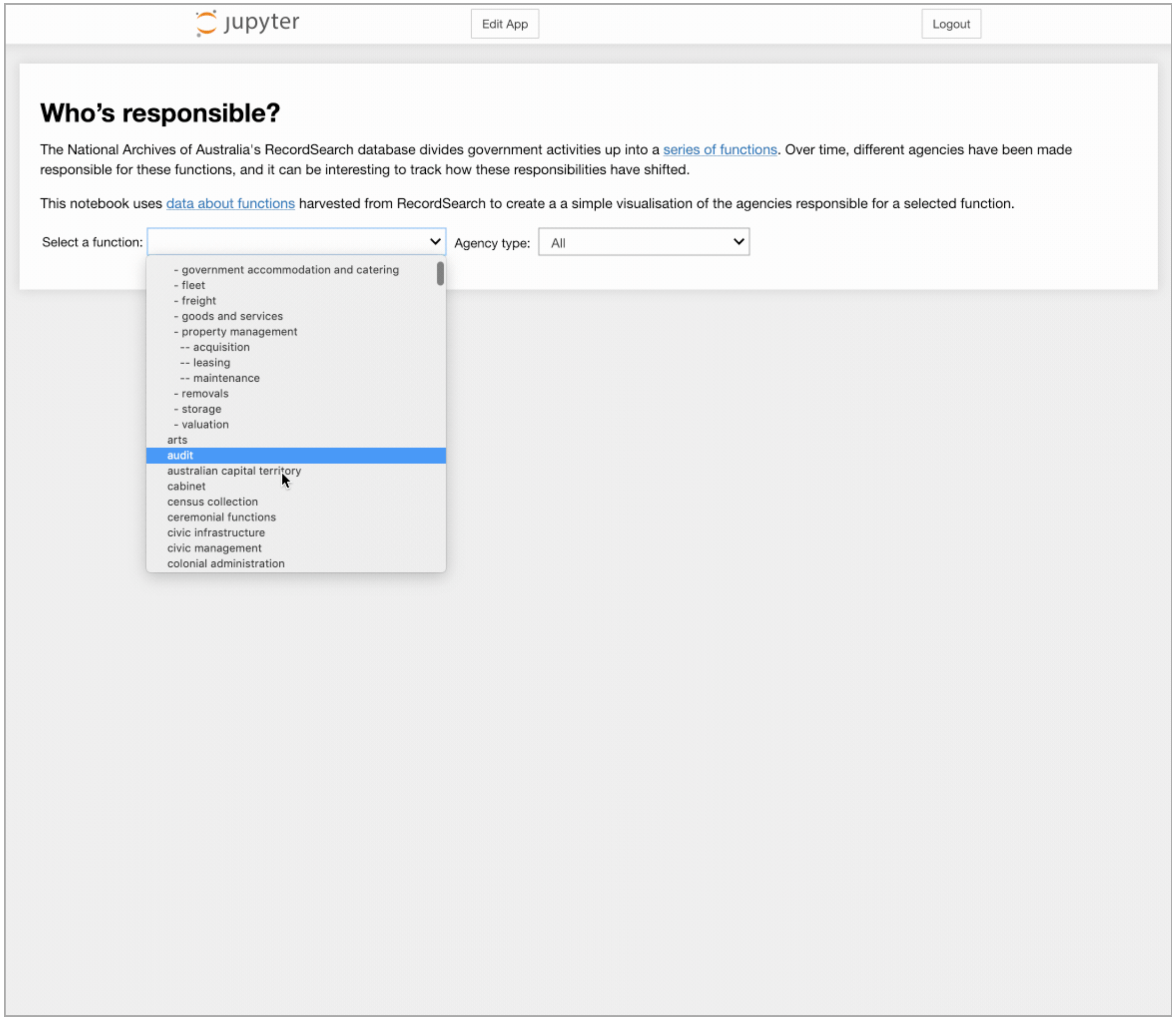
Structured but not downloadable



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/naa-wap/

Structured but not downloadable



Who’s responsible?

The National Archives of Australia's RecordSearch database divides government activities up into a series of functions. Over time, different agencies have been made responsible for these functions, and it can be interesting to track how these responsibilities have shifted.

This notebook uses data about functions harvested from RecordSearch to create a a simple visualisation of the agencies responsible for a selected function.

Select a function:

Agency type: All

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

Structured but not downloadable

https://glam-workbench.github.io/nsw-state-archives/

Downloadable but in many separate pieces

https://player.vimeo.com/video/321657685?api=1

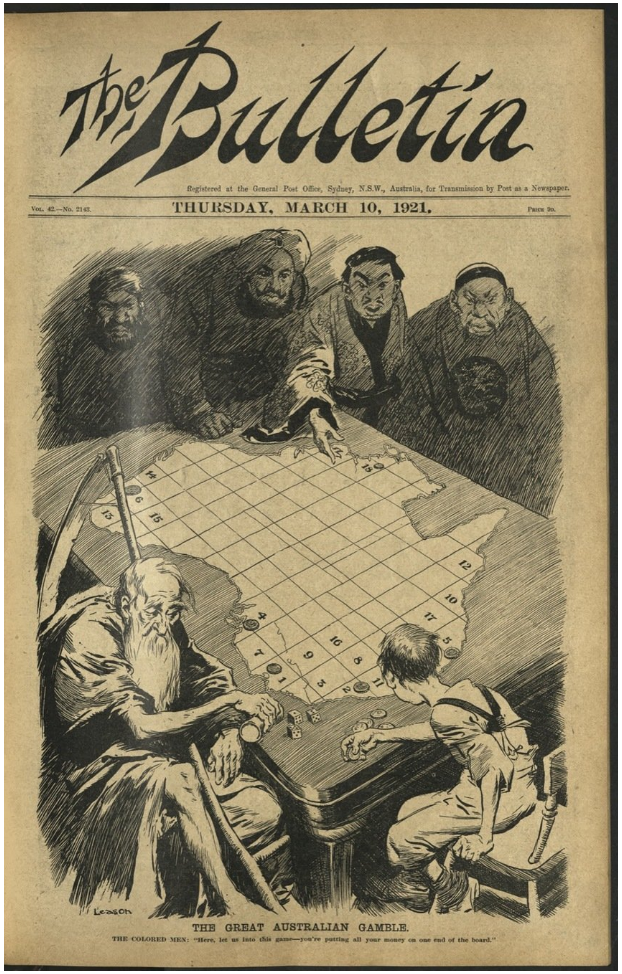
https://glam-workbench.github.io/hansard/

Downloadable but in many separate pieces

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

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

Downloadable but in many separate pieces



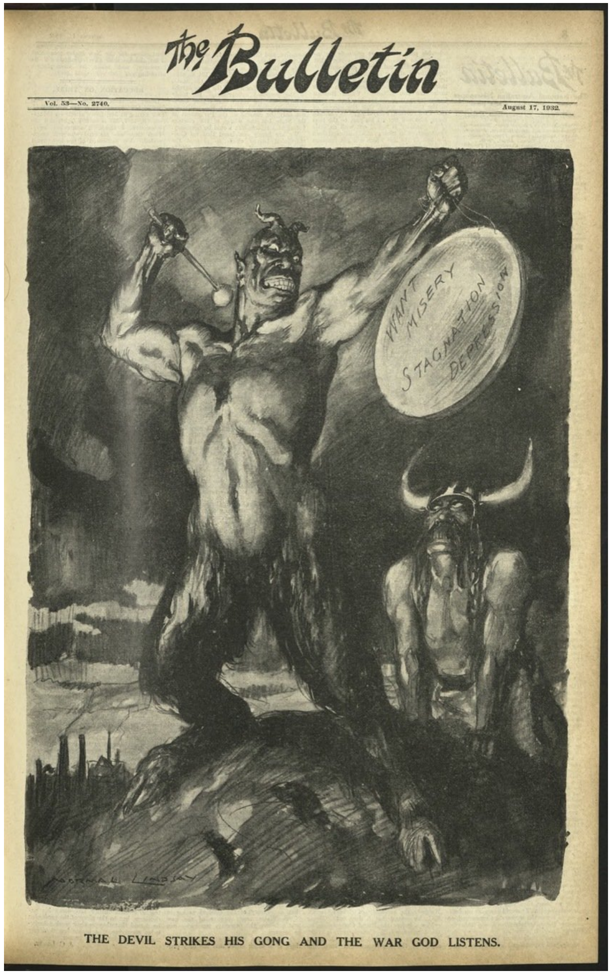
The Bulletin

Registered at the General Post Office, Sydney, N.S.W., Australia, for Transmission by Post as a Newspaper.

Vol. 42.—No. 2143. THURSDAY, MARCH 10, 1921. Price 9d.

THE GREAT AUSTRALIAN GAMBLE.

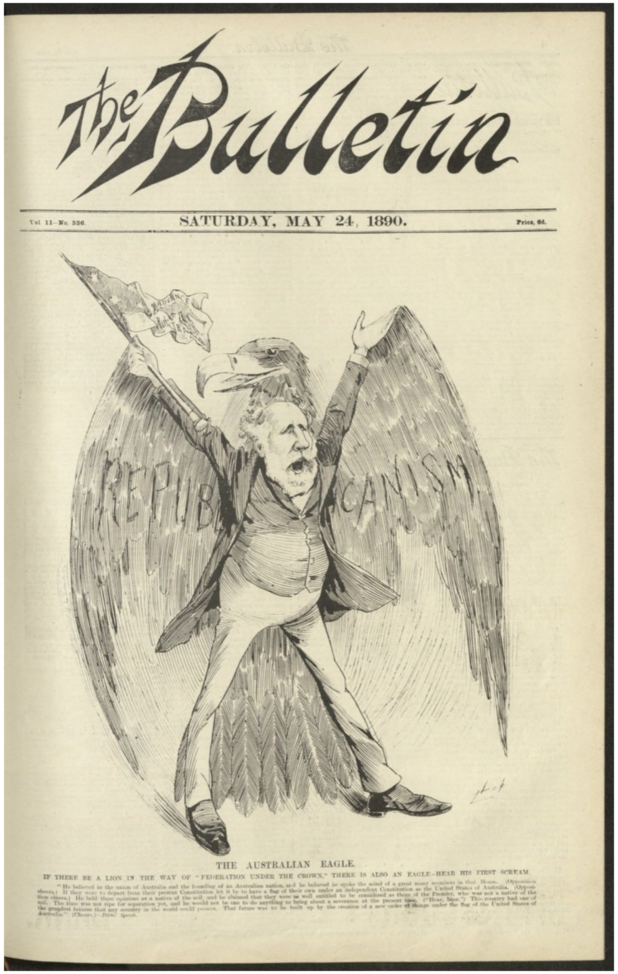
THE COLORED MEN: "Here, let us into this game—you're putting all your money on one end of the board."



The Bulletin

Vol. 53—No. 2740. August 17, 1932.

THE DEVIL STRIKES HIS GONG AND THE WAR GOD LISTENS.



The Bulletin

Vol 11—No. 536. SATURDAY, MAY 24, 1890. Price, 6d.

THE AUSTRALIAN EAGLE.

IF THERE BE A LION IN THE WAY OF "FEDERATION UNDER THE CROWN," THERE IS ALSO AN EAGLE—HEAR HIS FIRST SCREAM.

"He believed in the union of Australia and the founding of an Australian nation, and he believed he spoke the mind of a great many members in that House. (Opposition cheers.) If they were to depart from their present Constitution let it be to have a flag of their own under an independent Constitution as the United States of Australia. (Opposition cheers.) He held these opinions as a native of the soil and he claimed that they were as well entitled to be considered as those of the Premier, who was not a native of the soil. The time was not ripe for separation yet, and he would not be one to do anything to bring about a severance at the present time. ("Hear, hear.") This country had one of the grandest futures that any country in the world could possess. That future was to be built up by the creation of a new order of things under the flag of the United States of Australia." (Cheers.)—*Dibbs' Speech.*

https://updates.timsherratt.org/2019/05/09/over-the-last.html https://timsherratt.org/

https://timsherratt.org/