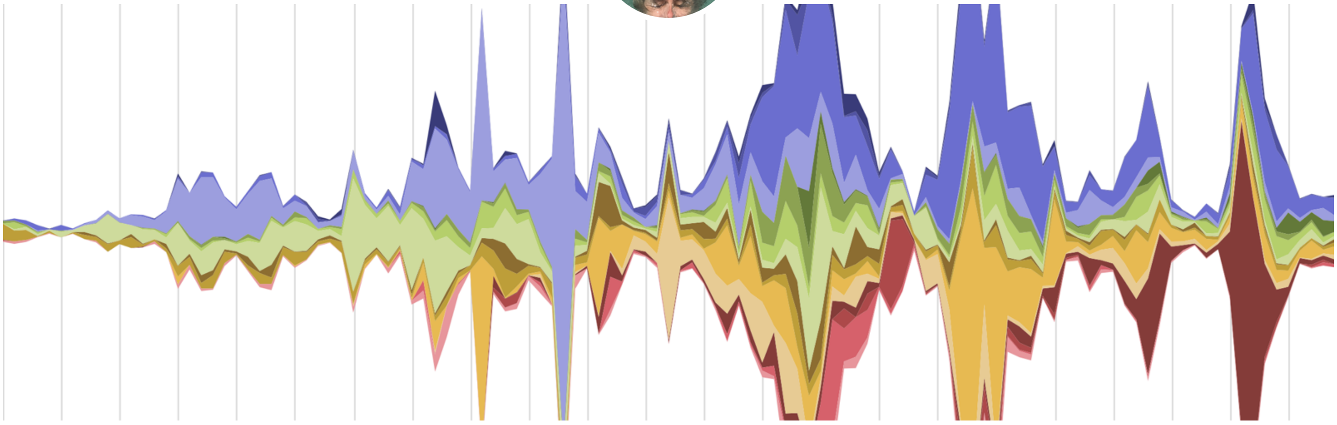


Who belongs? Reading identity, ownership, and legitimacy

by Tim Sherratt

14 February 2019

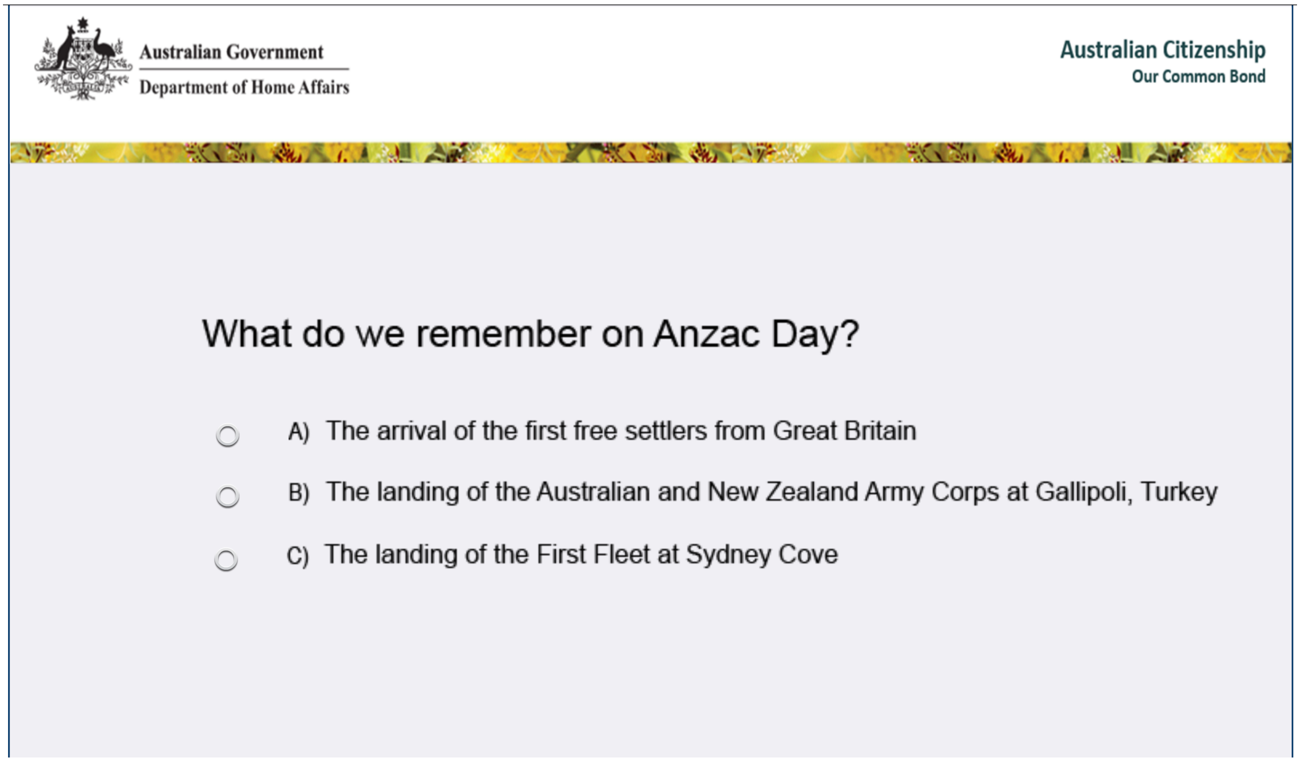
CATEGORIES research

Invited presention at From text to data – new ways of reading, National Library of Sweden, 5 February 2019.

The full set of slides are also available, as is all the code and data.

Let's start with a question you've probably always wondered about: 'Could I be an Australian?'.

Let's see… Here's a page from the Australian Citizenship Test. Applicants for citizenship need to answer 15 out of 20 multiple choice questions correctly.



Australian Government
Department of Home Affairs

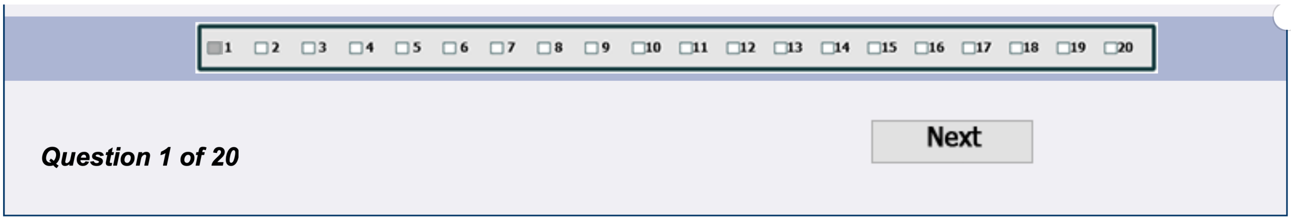
Australian Citizenship
Our Common Bond

What do we remember on Anzac Day?

○ A) The arrival of the first free settlers from Great Britain

○ B) The landing of the Australian and New Zealand Army Corps at Gallipoli, Turkey

○ C) The landing of the First Fleet at Sydney Cove



Try the Practice Test online

But apparently this isn't hard enough. In 2017 the government announced it would add a tougher English language requirement, as well as new questions that would test applicants' understanding of 'Australian values'.

The Prime Minister was unable at the time to clearly explain what these 'Australian values' were. Fortunately Twitter was there to help. The #australianvalues hashtag was soon trending as the public offered a range of suggestions of what it meant to be Australian.

Being a good digital historian, I used the Documenting the Now project's Twarc tool to document the resulting tweet storm.

Here's a few of the suggestions:

- Saying "is it hot enough for you?" as a form of greeting in summer. #AustralianValues
- Never fucking ever barrack for New Zealand in any sport #AustralianValues
- Being used as imperialist cannon-fodder. #AustralianValues
- Finger wave when passing each other on long country highway; xenophobia #AustralianValues
- Casual racism #AustralianValues

The government's proposed changes, which have yet to pass through parliament, are just one example of the way they have sought to portray citizenship as key weapon in the defence of national security. As well as making it harder to get, they've also introduced changes that would make it easier to take it away from those convicted of terrorism-related offences. As an election approaches in 2019, debates around 'border security' will continue. As the #australianvalues debacle illustrates, the borders in need of protection are not only the continent's geographic boundaries, but those that define who belongs.

Somewhat ironically, a number of Australian politicians have discovered in the last couple of years that they don't belong in parliament. Barnaby Joyce, our deputy prime minister, for example, found out he was a New Zealander. Unknowingly, he had gained NZ citizenship through his father, who was born there, and this ran afoul of the Australian Constitution's requirement that members of parliament should not be 'a subject or a citizen or entitled to the rights or privileges of of a subject or a citizen of a foreign power'.

Borders shift. Until 1948 there was no such thing as an Australian citizen — Australians were just British subjects. The deputy PM's ties to New Zealand would have been no impediment to sitting in parliament, as New Zealanders were also British subjects. Over the twentieth century, Australia's imperial brethren gradually became 'foreign'.

In this talk today I want to look at some changes in the language of belonging in Australia, focussing on two words, 'aliens' and 'immigrants'.

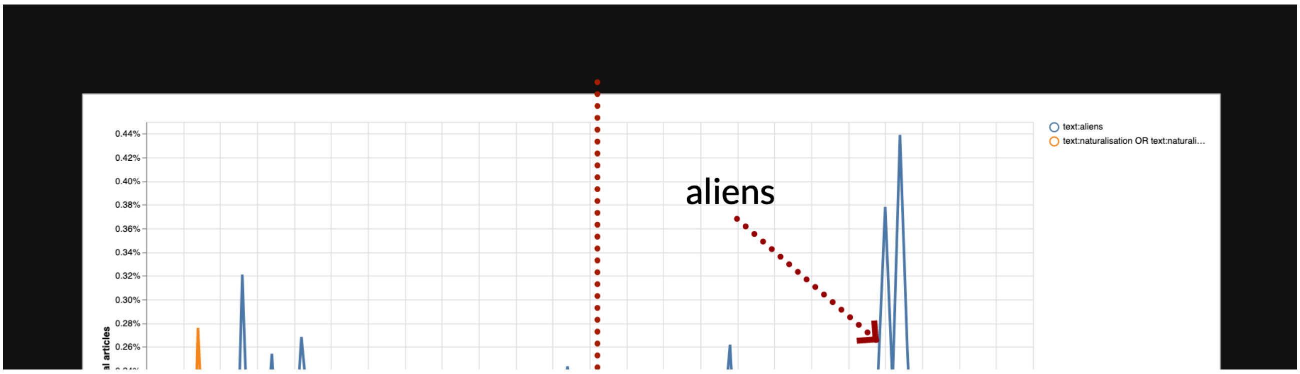
I'm not going to present a detailed argument, nor am I going to offer a very sophisticated analysis. I'm mostly just counting words and pointing to interesting things. But I suppose even simple techniques can be pretty remarkable when they're applied across large historical collections.

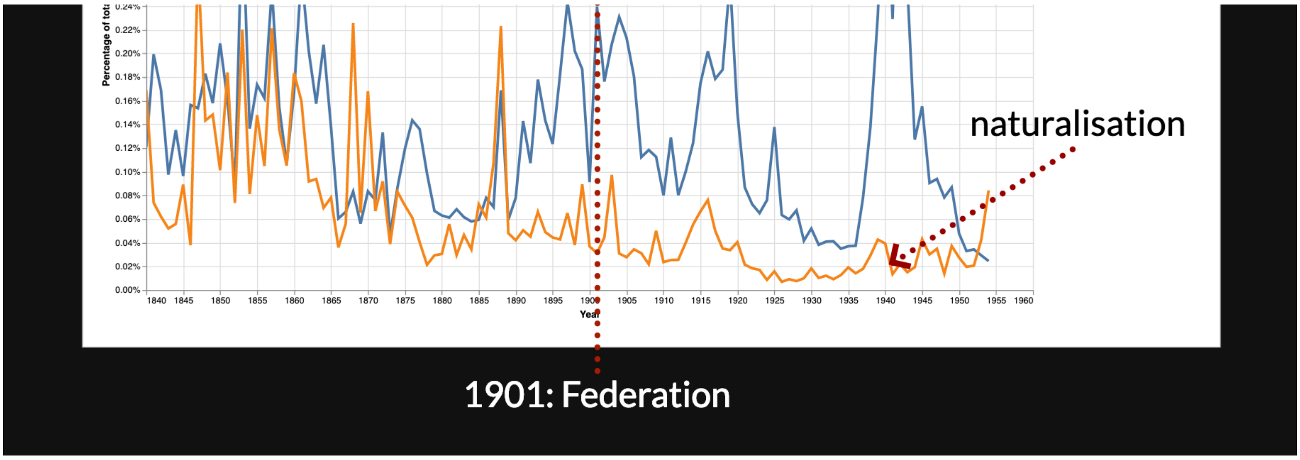
I also want to think about how we get the texts to the tools. I spend a lot of time hacking at the plumbing of digital collections — making connections, and unclogging pipes — trying to get the data to flow. While institutions digitise terabytes of text it's often still a struggle to get the data out in a form that can be easily used and shared. Stories about moving data around are not very glamorous, but they're important in understanding the limits of our infrastructures.

Let's start with 200 million newspaper articles, documenting a large part of Australia's post-invasion history. Trove's digitised newspapers are a wonder, but they're also overwhelming. What does it mean when you type in a query and get back 5 million results?

Trove's API delivers full text and metadata from the digitised newspapers in a form computers can understand and manipulate. But it also provides summary data, or facets, that we can use to create quick pictures of our searches — how do those millions of results break down year by year, for example?

Here we're comparing the occurrence of two words over time — 'naturalisation' and 'aliens'. Instead of the raw number of results, the chart is showing the matches as a proportion of the total articles for each year.



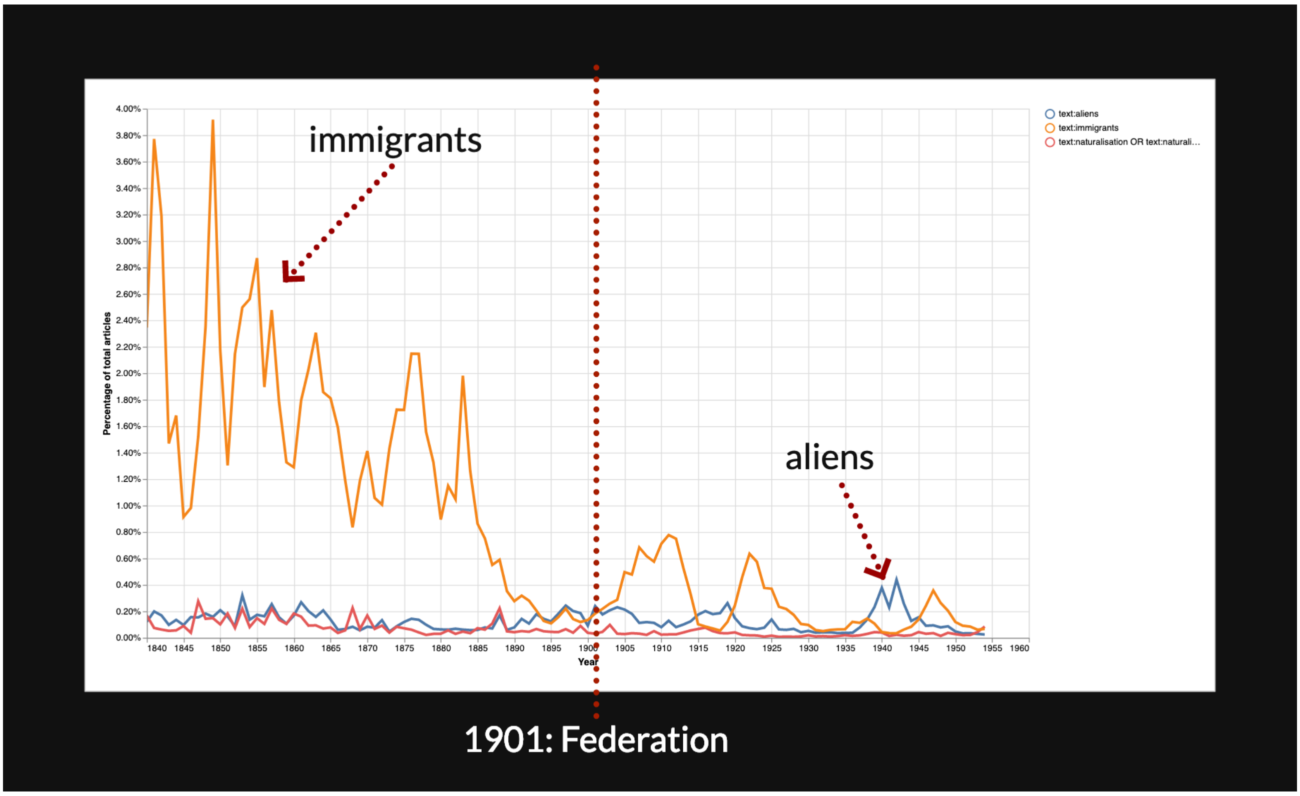


The code I used to create this chart is available as a Jupyter notebook. It's the latest version of a tool called QueryPic that I first created many years ago. Like the other examples in my GLAM Workbench, you can run the QueryPic code live in your browser, create your own charts, and download the results. The workbench is intended to encourage experimentation and learning by providing useful tools and examples that expose their own workings. Let me know if you'd like a demo!

I'm comparing 'aliens' and 'naturalisation' because their meanings are historically entwined. An alien is someone from a foreign country. Naturalisation is the process by which a resident alien swears allegiance to their new home and is granted the rights that come with citizenship. The first half of the nineteenth century looks a bit messy, but you can see that the two terms follow each other fairly closely. This was a period when the Australian colonies passed various pieces of legislation relating to the naturalisation of non-British settlers. In South Australia, for example, there was a large population of German immigrants for whom naturalisation offered greater rights to land ownership.

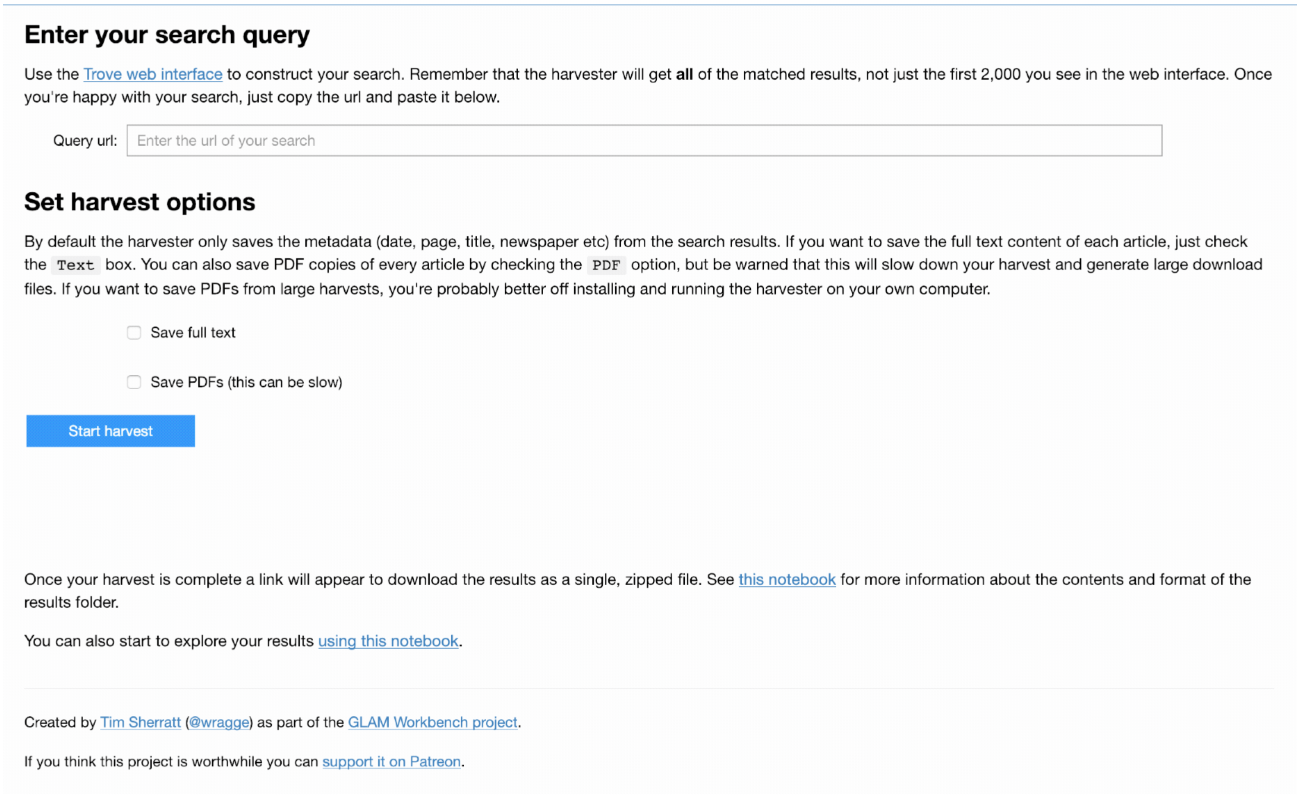
But in the 1880s and 90s the terms start to diverge. This suggests the word 'aliens' was being used in a broader range of contexts.

If we throw the word 'immigrants' into the mix, we see that its occurrence drops away in the same period until it's roughly on par with 'aliens'. Setting aside the brutal irony of British colonists describing others as 'aliens' in a land they'd stolen from its original inhabitants, it seems that something unusual was happening in the 1890s.



Indeed it was. The 1890s are notable in Australian history for a number of reasons. Economically, the eastern colonies plunged into depression, bringing an end to a period of rapid growth powered by gold and wool. Industrial conflict pushed trade unions into the political arena with the formation of the Labor Party. Nationalism was on the rise both culturally and politically, as the separate colonies moved towards federation in 1901. And attempts by the colonies to restrict Chinese immigration were transformed into a vision of national identity — the new nation of Australia would be young, strong, and 'white'. In the early twentieth century this vision took legislative form as the White Australia policy.

Let's dig a bit deeper. The Trove API can also be used to harvest large quantities of digitised newspaper articles, included the OCRd text. That's great if you know how to code. For those who don't I created a Trove Newspaper Harvester. You feed it the url of a Trove search, and it gives you back a CSV file with metadata of all the matching articles, as well as the contents of each article in a separate text file. The original harvester is a Python command-line tool, but to lower the barriers that bit more I've recently created an app-ified version in a Jupyter notebook — just cut, paste, point and click.



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.

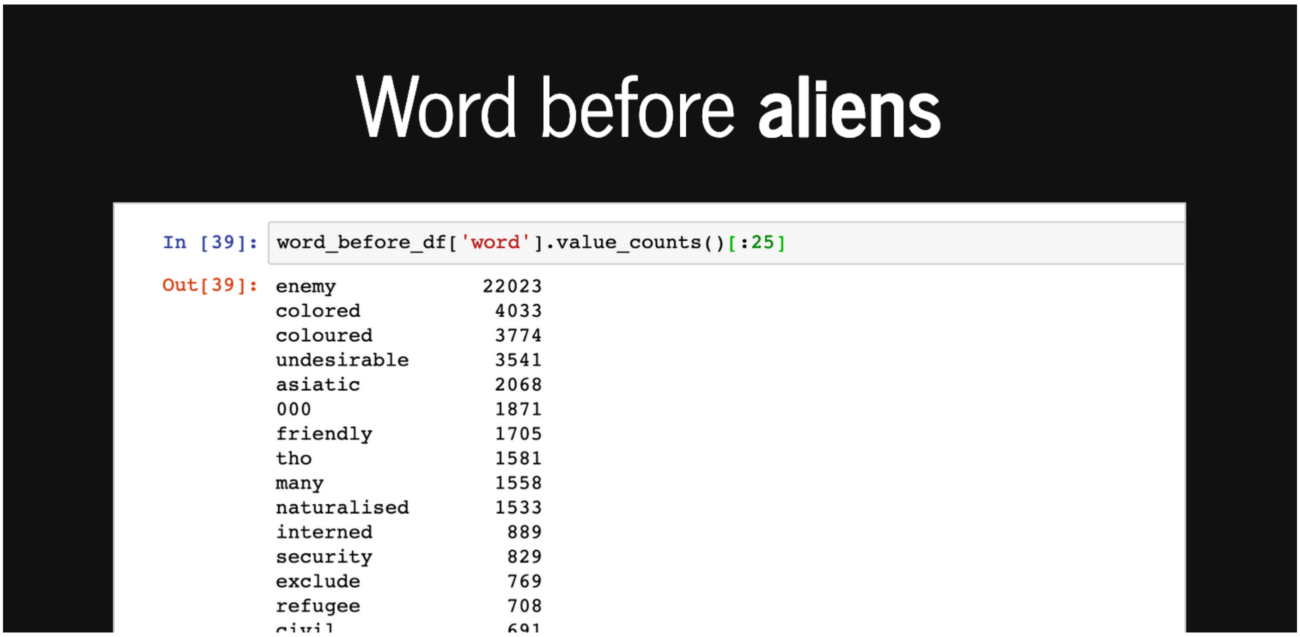
Try it live on Binder

If you browse around my GLAM Workbench you'll see lots of little 'Use live on Binder' buttons. These spin up the notebooks in a customised computing environment using a fabulous project called Binder. So it's entirely possible to harvest many thousands of newspaper articles from Trove using nothing but your browser. Obviously for really big, or frequent, harvests you'd want to set up your own environment, with Jupyter installed locally. For example, I recently used the harvester notebook to download more than 2 million articles that included the word 'Chinese'. It took about 13 hours.

For today's experiment I harvested two batches of newspaper articles: 180,000 articles containing the word 'aliens', and half a million articles that included 'immigrants'. I then ran a script across all the little text files to locate the target word within the text and grab five words on either side. I saved the results in a couple of CSV files for further examination.

You've probably already guessed that I've shared the code for all this in a series of Jupyter notebooks. You could easily adapt the code to explore the occurrence of other words.

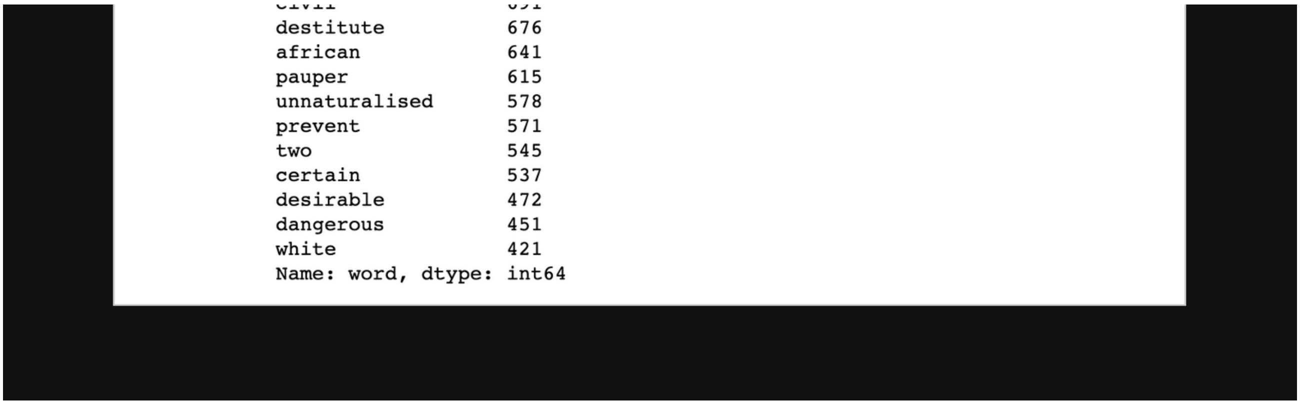
Let's look at the words that appeared most frequently before 'aliens' in Australian newspapers, excluding standard stop words. 'Enemy' was by far the most common, but 'coloured', 'undesirable', and 'asiatic' are also near the top of the list.



Word before **aliens**

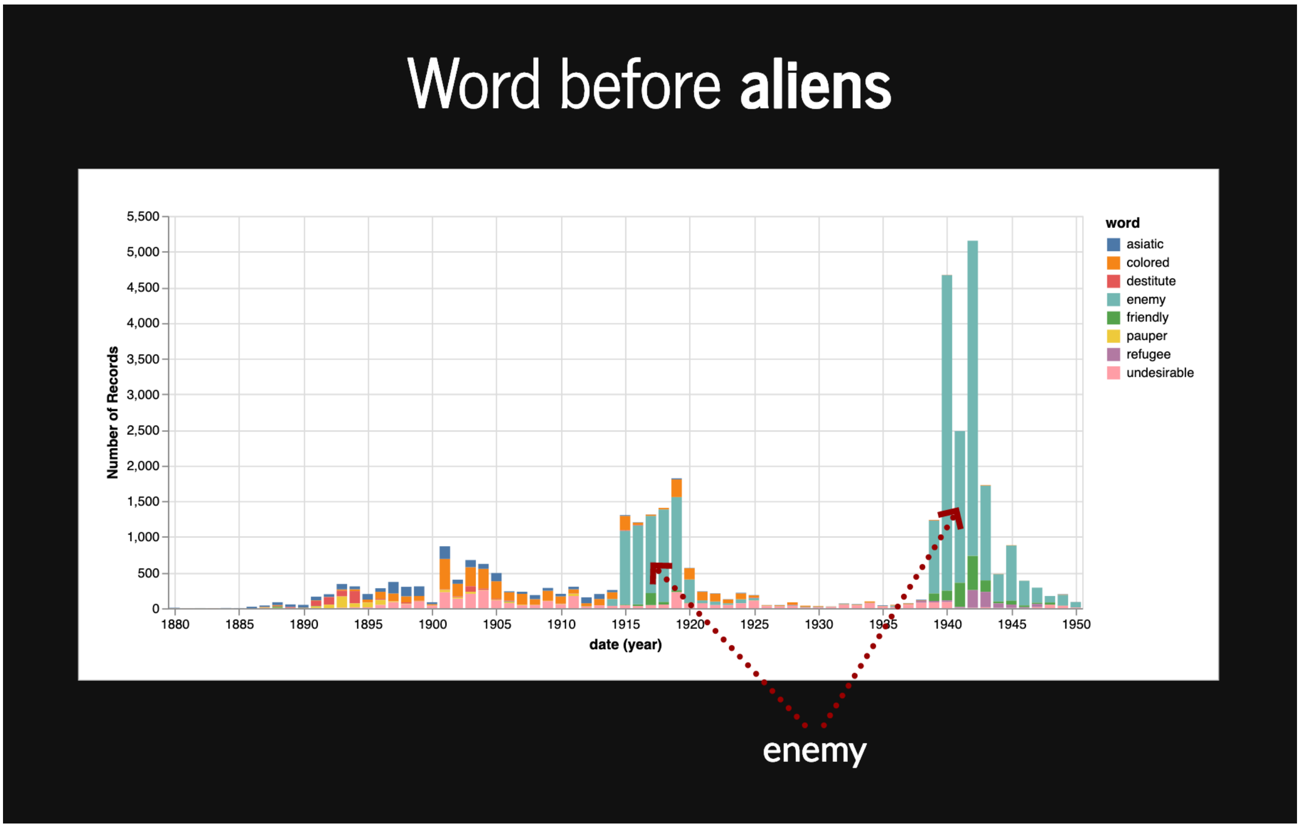
```
In [39]: word_before_df['word'].value_counts()[:25]

Out[39]: enemy           22023
         colored          4033
         coloured         3774
         undesirable      3541
         asiatic          2068
         000              1871
         friendly         1705
         tho              1581
         many             1558
         naturalised      1533
         interned          889
         security          829
         exclude           769
         refugee           708
         civil             691
```



```
civil               691
destitute           676
african             641
pauper              615
unnaturalised       578
prevent             571
two                 545
certain             537
desirable           472
dangerous           451
white               421
Name: word, dtype: int64
```

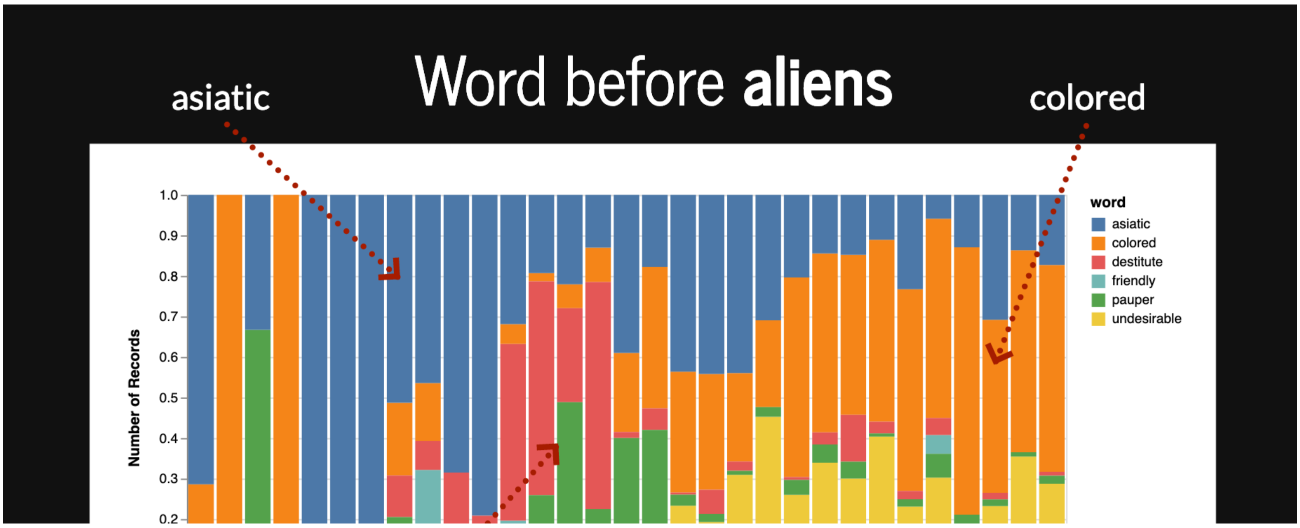
By plotting the frequencies over time, we can see, not surprisingly, that use of the term ‘enemy aliens’ peaks during the two world wars.



Word before **aliens**

enemy

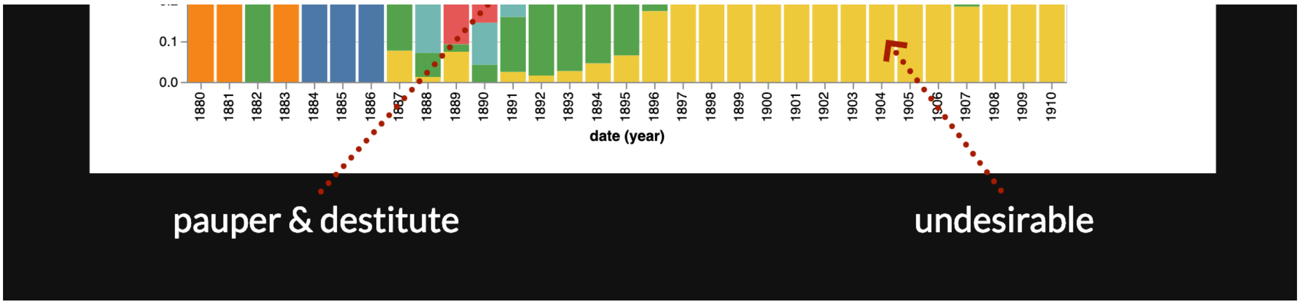
But for now I want to focus on the period before the first world war where we observed some of those large-scale changes. Here’s the same selection of words from 1880 to 1910, displayed in their relative proportions. It’s interesting that ‘pauper’ and ‘destitute’ are prominent at a time of high unemployment and industrial unrest. Dipping back into the articles you’ll see that the immigration of ‘pauper aliens’ was identified as cause for concern across the British empire, but the language was clearly picked up in Australia where an ‘Anti Pauper Aliens League’ was established.



asiatic

Word before **aliens**

colored



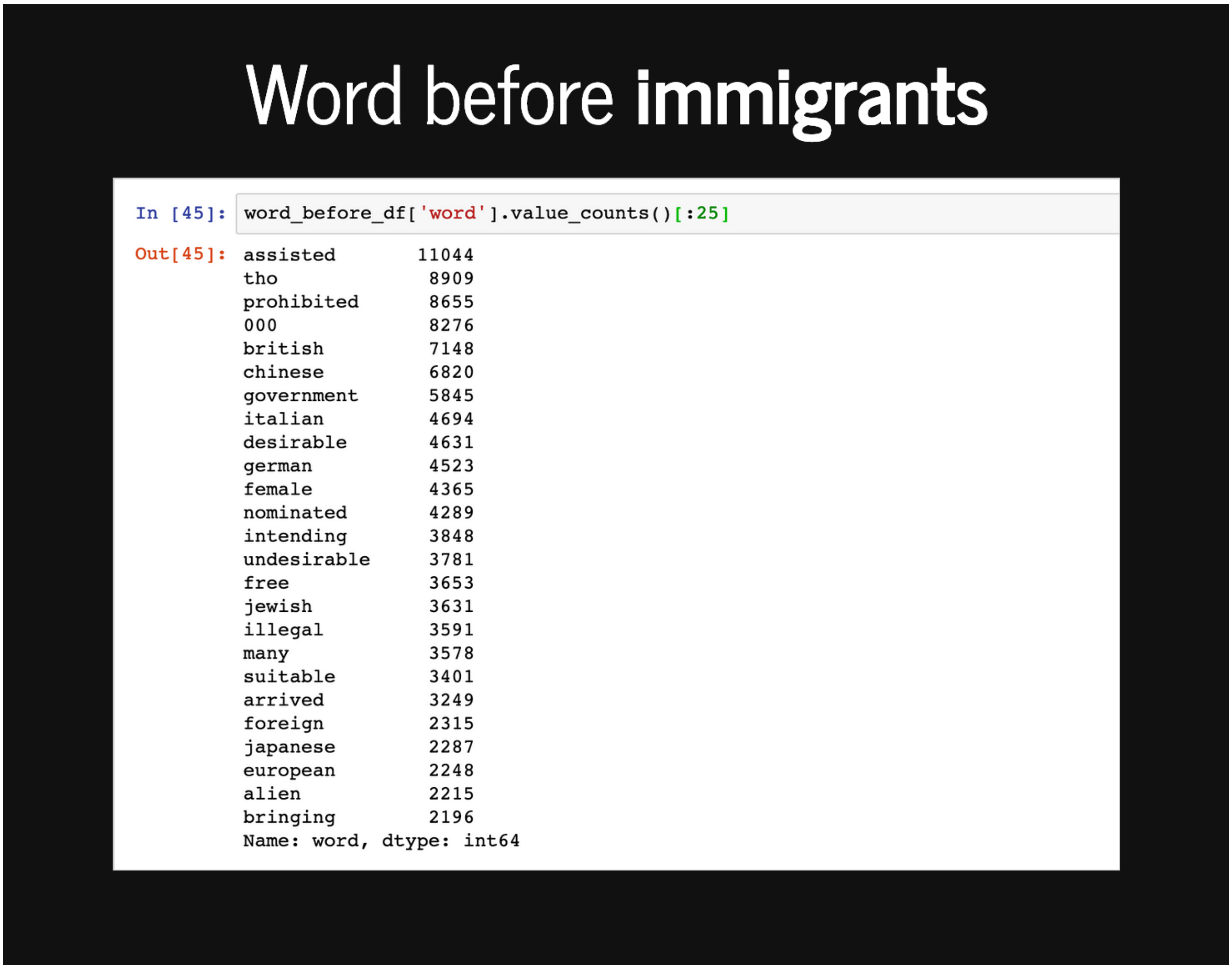
pauper & destitute

undesirable

Does the shift from ‘asiatic’ to ‘coloured’ suggest something about the broadening of racial categories? Certainly this was occurring at a political level, where legislation to restrict Chinese arrivals was expanded to catch all non-white immigrants.

Both of these changes seem to fit with the appearance of ‘undesirable’ as a category of aliens. Not all foreign arrivals were equal. Not all belonged in Australia.

Similar patterns appear in words that occur before ‘immigrants’.

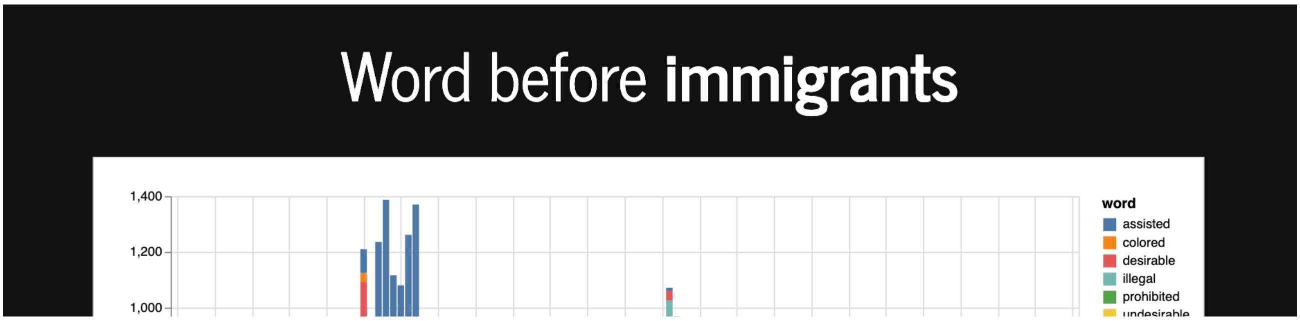


Word before **immigrants**

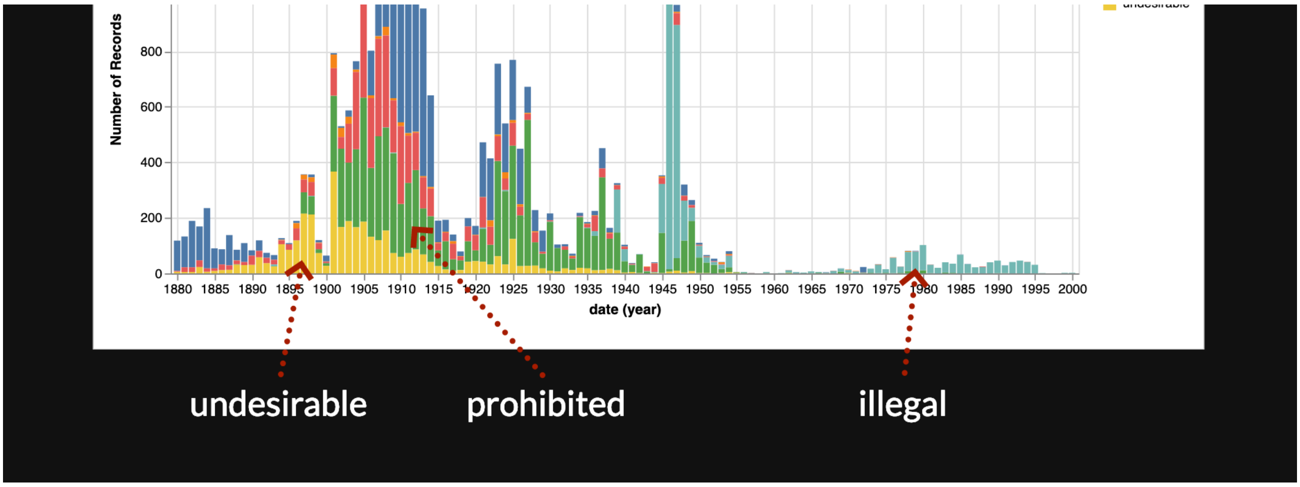
```
In [45]: word_before_df['word'].value_counts()[:25]

Out[45]: assisted       11044
         tho             8909
         prohibited      8655
         000             8276
         british         7148
         chinese         6820
         government      5845
         italian         4694
         desirable       4631
         german          4523
         female          4365
         nominated       4289
         intending       3848
         undesirable     3781
         free            3653
         jewish          3631
         illegal         3591
         many            3578
         suitable        3401
         arrived         3249
         foreign         2315
         japanese        2287
         european        2248
         alien           2215
         bringing        2196
         Name: word, dtype: int64
```

Both ‘undesirable’ and ‘desirable’ peak in the late nineteenth and early twentieth centuries. In 1901, the Immigration Restriction Act, the cornerstone of the White Australia Policy, gave legal force to these sorts of distinctions and created the category of ‘prohibited’ immigrants to exclude the unwanted. Although most of Trove’s digitised newspapers were published before 1955 for copyright reasons, there are enough from the later period for to suggest a transition from ‘prohibited’ to ‘illegal’.

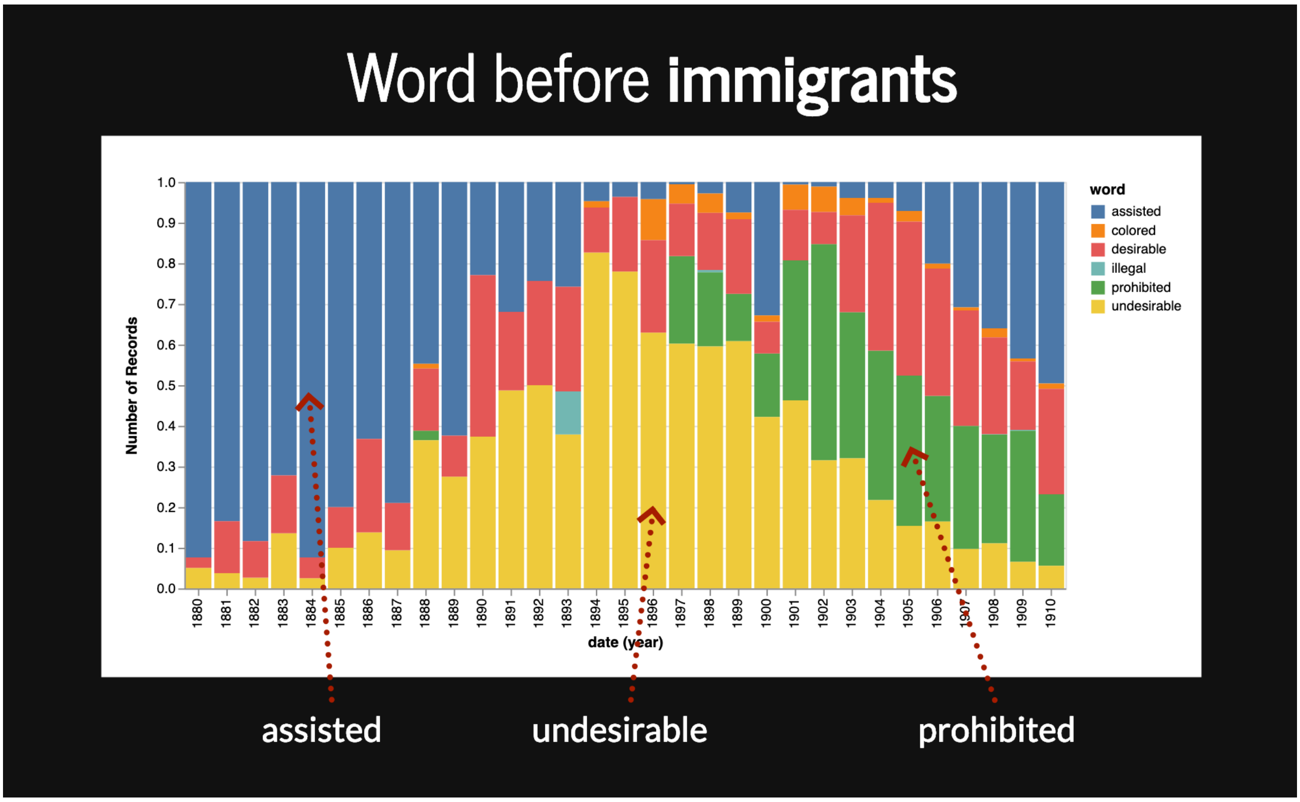


Word before **immigrants**



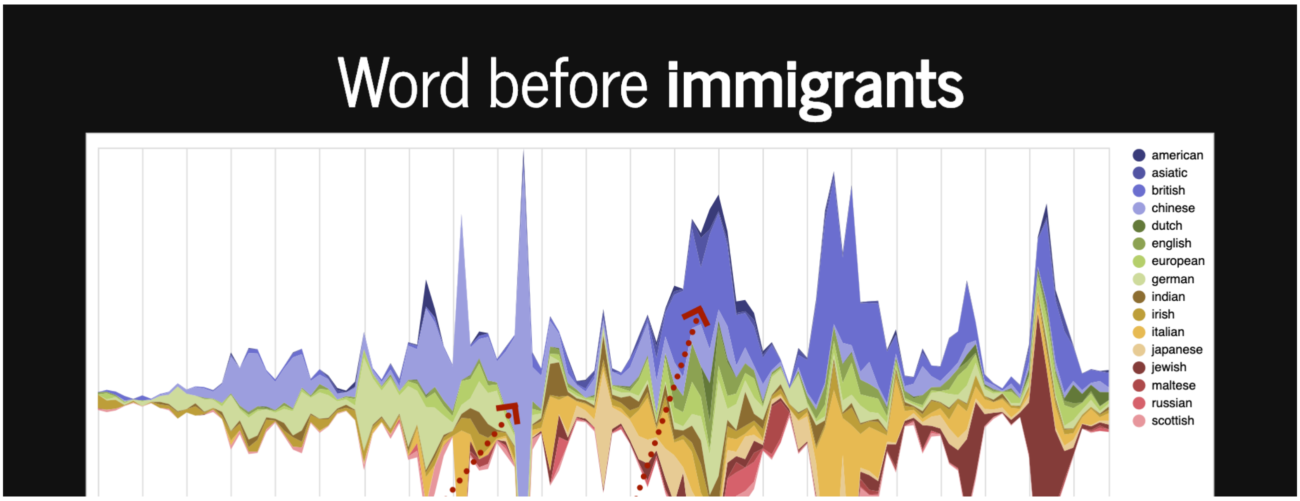
undesirable prohibited illegal

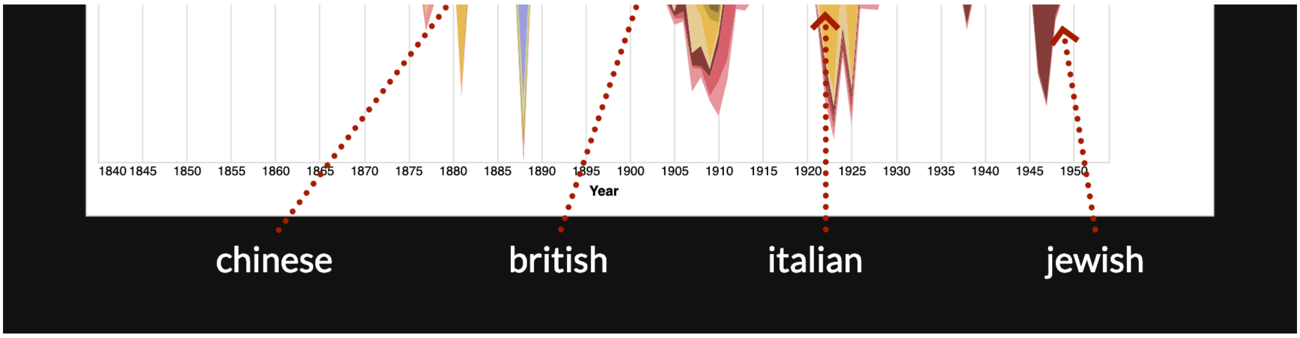
Focusing again on the period around 1901, there seems to be a shift from discussion relating to the encouragement of immigration, reflected in the use of the term ‘assisted’, to issues of selection and control. As historians have noted, security played a greater role in shaping immigration policy after Federation. If Australia was to be ‘white’, it needed the right sort of settlers to occupy and defend its vast spaces.


Word before **immigrants**

assisted undesirable prohibited

Also amongst the words appearing before ‘immigrants’ are ethnic and national groups. I thought these would also be interesting to observe through time.


Word before **immigrants**



chinese british italian jewish

Of course the peaks here say nothing about the number of immigrants arriving or living in Australia — it's just a measure of how much they were being talked about. So the spike in 'chinese' immigrants in the 1880s reflects the introduction of anti-Chinese legislation. In the 1920s, concerns that some southern Europeans were not quite 'white' enough, no doubt contributed to the increased frequency of 'italian'.

Let's take this exploration further. *The Bulletin* was the most widely-read weekly newspaper in Australia, and fostered a brand of radical nationalism. In the early twentieth century, it's masthead included the slogan 'Australia for the White Man'.

The Bulletin from 1880 to 1968 has been digitised and is available on Trove. But not in the newspapers zone. Bulletin articles pop up in the 'Journals' zone, as do a growing range of full-text digitised magazines, journals, periodicals, and newsletters. These resources are in a different backend system from the newspapers. So while you can manually download the OCRd text of each issue of the *The Bulletin* from the web interface, you can't get it through the Trove API.

Fortunately, I was able to scrape a list of issue identifiers from the journal's browse interface, and use them to automate the download of all the text files. Once again, there's a Jupyter notebook with all the code and explanation, but if you just want the texts, you can grab all 4,534 of them from this GitHub repository.

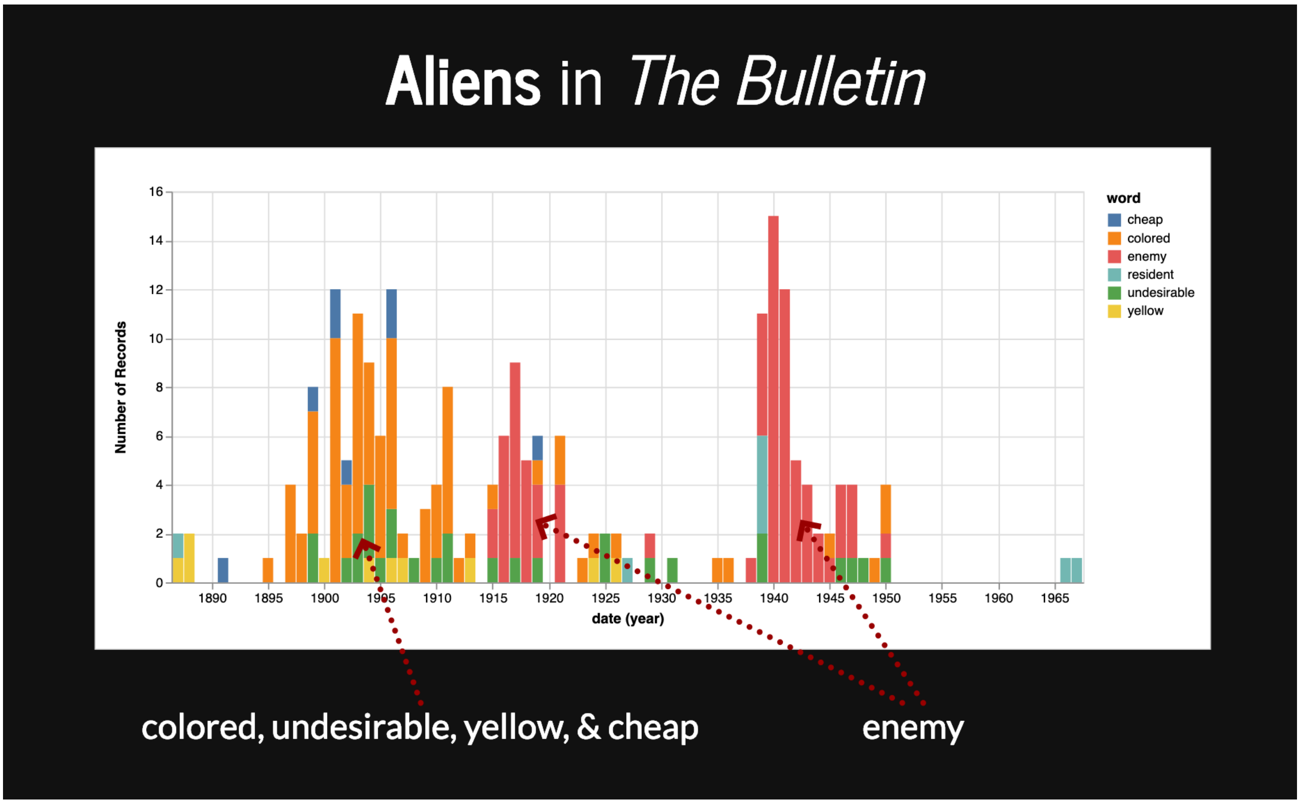
Another important collection of digitised texts are the Commonwealth Parliamentary Debates, also known as Hansard. All speeches made in the Australian parliament from 1901 have been digitised by the Parliamentary Library and marked up as well-structured XML files. It's a fabulous resource, which unfortunately is hidden behind a horrible search interface.

I've harvested all the XML files from their interface and made them available through a GitHub repository. (There's no notebook for that one yet…)

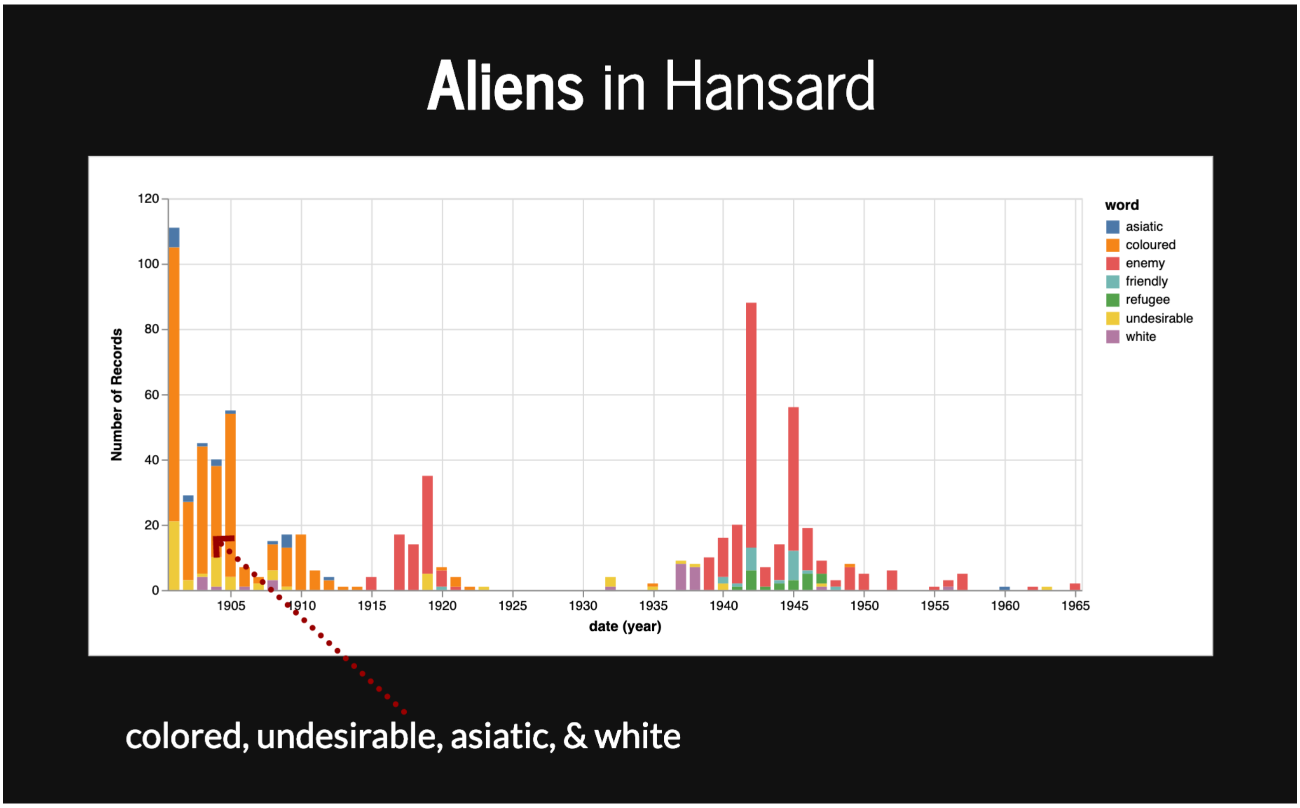
Along the way I also ended up building my own online version of Hansard, complete with a new search interface. (It's a long story…) One feature of my search interface is that you can save your results as a CSV file — an easy first step in the exploration of text as data.

So as a result of all this scraping and hacking I happen to have all the text from *The Bulletin* and Hansard on my computer. Why not do something with it? I used much the same methodology as I had with the newspapers to pull out some data about words that appeared before 'aliens'.

The results are pretty familiar, Both show wartime peaks for 'enemy' aliens, and both point to discussions of race and 'suitability' in the early years of the twentieth century.


Aliens in *The Bulletin*

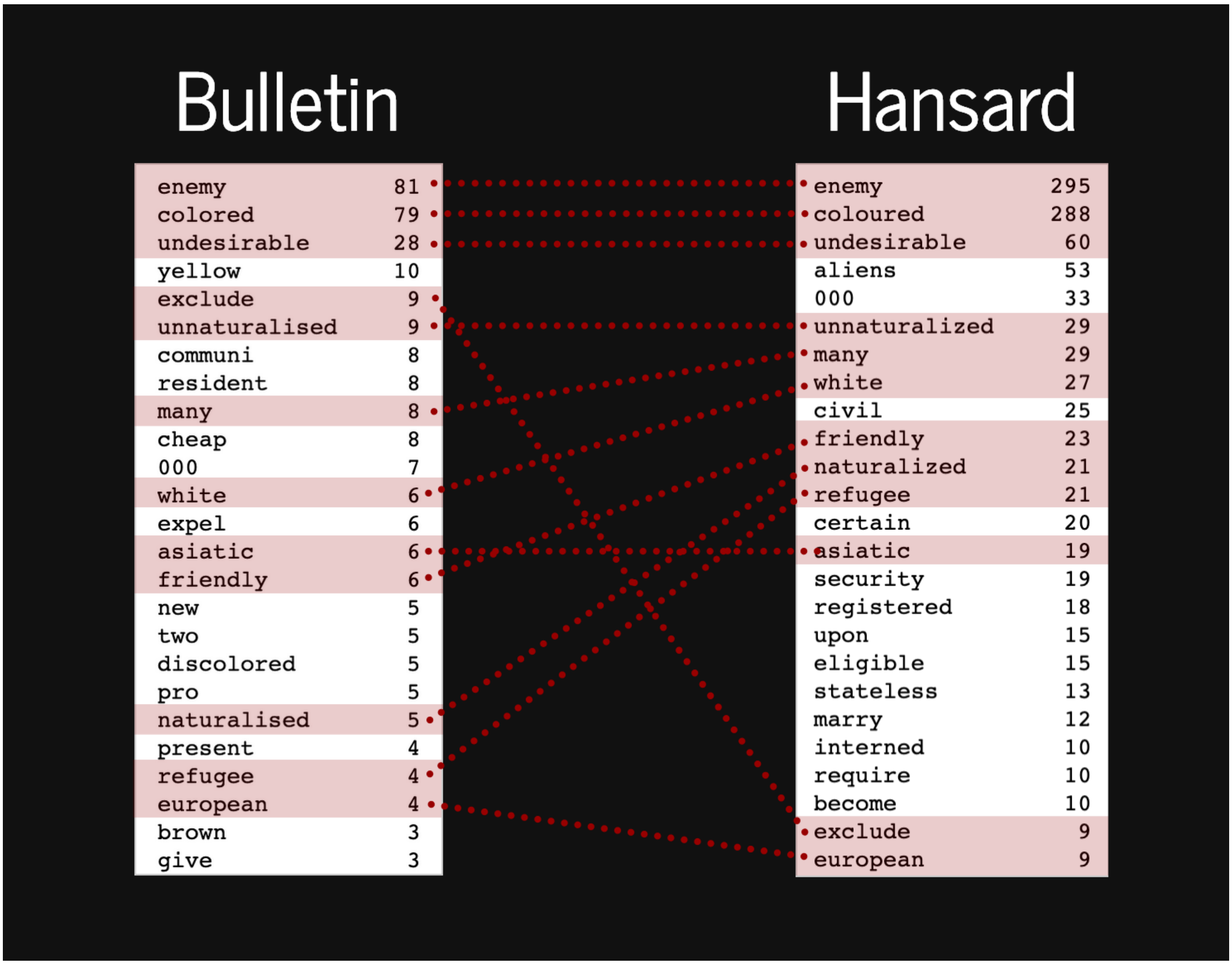
colored, undesirable, yellow, & cheap enemy



Aliens in Hansard

colored, undesirable, asiatic, & white

The Bulletin's language might be more explicitly racist, but I was actually surprised at how many of the top twenty-five words were the same.



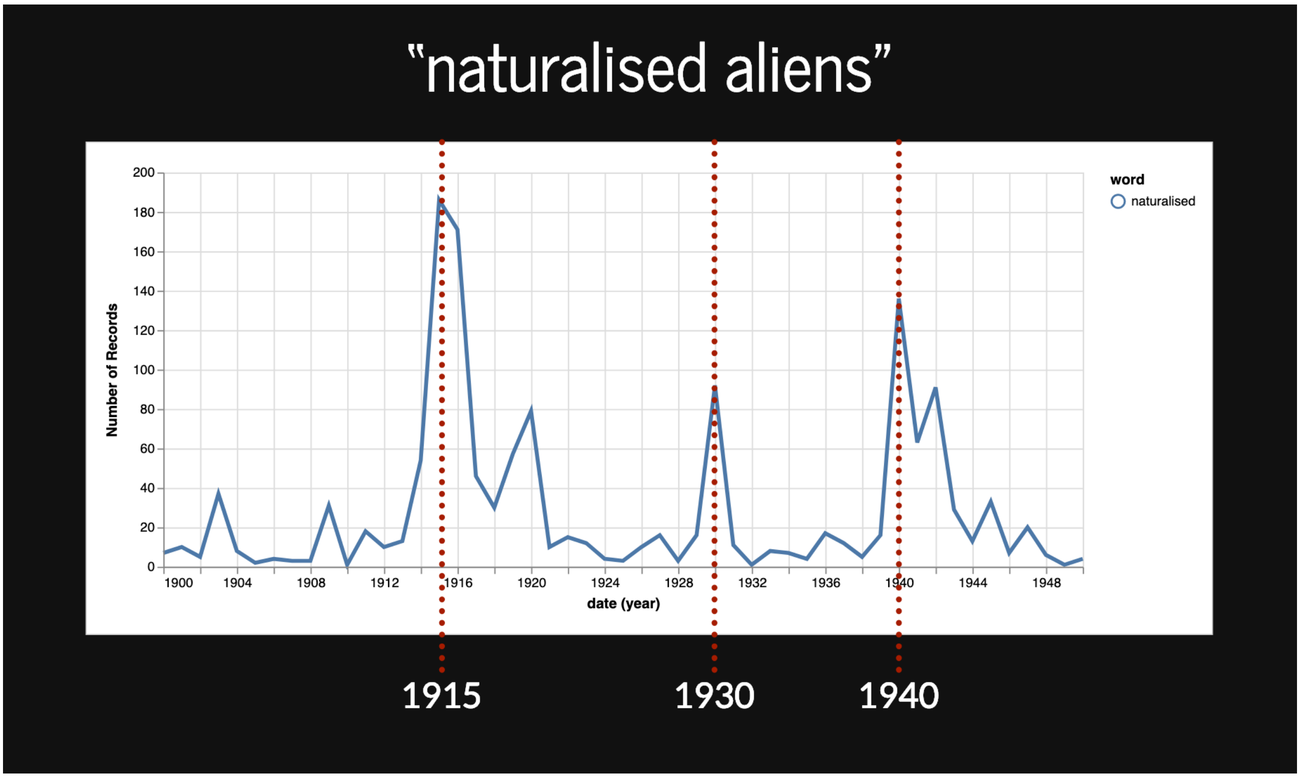
Bulletin	
enemy	81
colored	79
undesirable	28
yellow	10
exclude	9
unnaturalised	9
communi	8
resident	8
many	8
cheap	8
000	7
white	6
expel	6
asiatic	6
friendly	6
new	5
two	5
discolored	5
pro	5
naturalised	5
present	4
refugee	4
european	4
brown	3
give	3

Hansard	
enemy	295
coloured	288
undesirable	60
aliens	53
000	33
unnaturalized	29
many	29
white	27
civil	25
friendly	23
naturalized	21
refugee	21
certain	20
asiatic	19
security	19
registered	18
upon	15
eligible	15
stateless	13
marry	12
interned	10
require	10
become	10
exclude	9
european	9

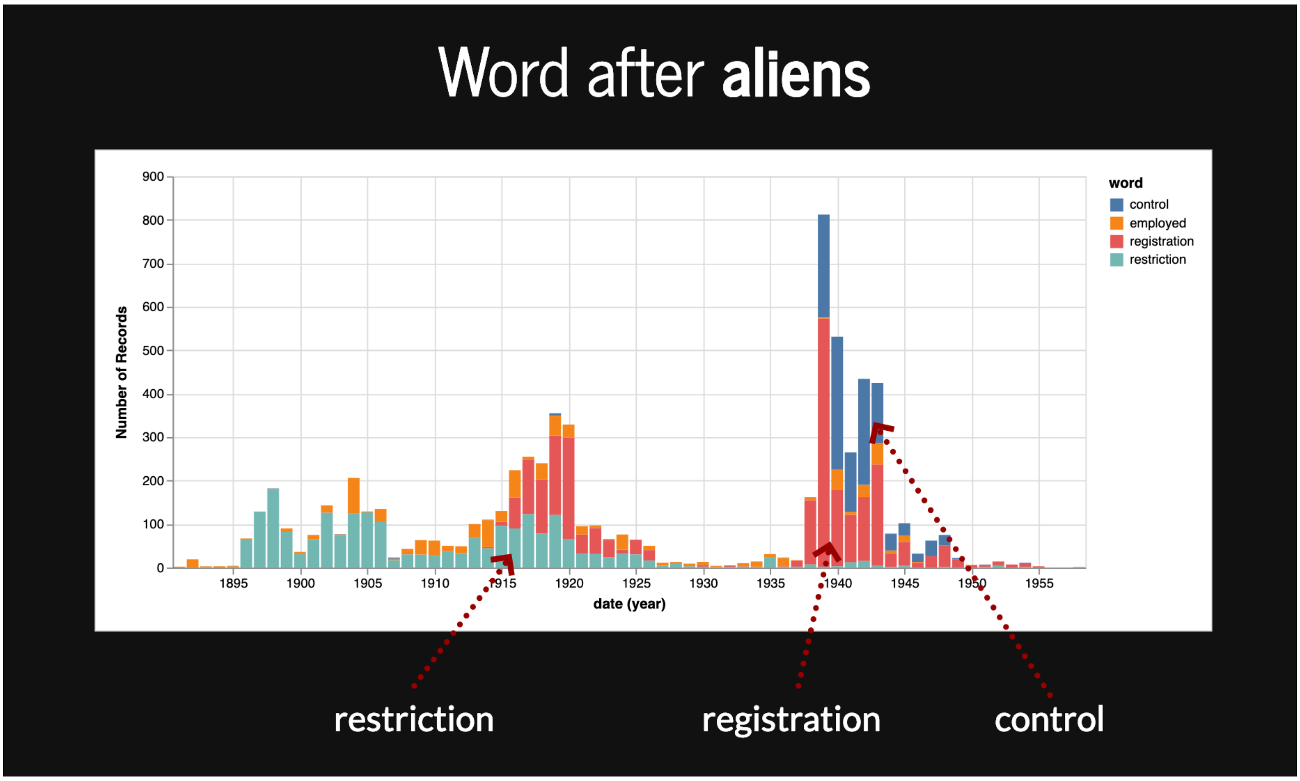
So, after all that — in a truly groundbreaking piece of digital research, I may have discovered that the White Australia Policy was racist… Ok, so there's nothing particularly unexpected in all these words, numbers and charts. But I think that it does underline the way language is mobilised and manipulated in the development and maintenance of racist systems like the White Australia Policy. Race was not mentioned in the Immigration Restriction Act. The mechanism of exclusion was the deceptively named Dictation Test, which could be given in any European language. Anyone who failed the test instantly became a 'prohibited immigrant' and could be arrested and deported. Words like 'undesirable' and 'illegal' help justify this sort of treatment.

But it was not only new arrivals whose status and legitimacy was subject to sudden change. What about all those 'enemy' aliens? What do you think happened to them? Another word that commonly occurs alongside 'aliens' is 'naturalised'. But surely if you're naturalised you're no longer an alien?

If we chart those occurrences we see, once again, that the peaks coincide with the wars. Being naturalised did not stop you being interned. Your allegiances remained suspect. You were alien first, Australian second. And the peak in 1930 seems to relate to the Prime Minister's insistence that federal departments should only employ naturalised aliens if 'British' workers were unavailable. 'Naturalised aliens' were British subjects, but not British enough…

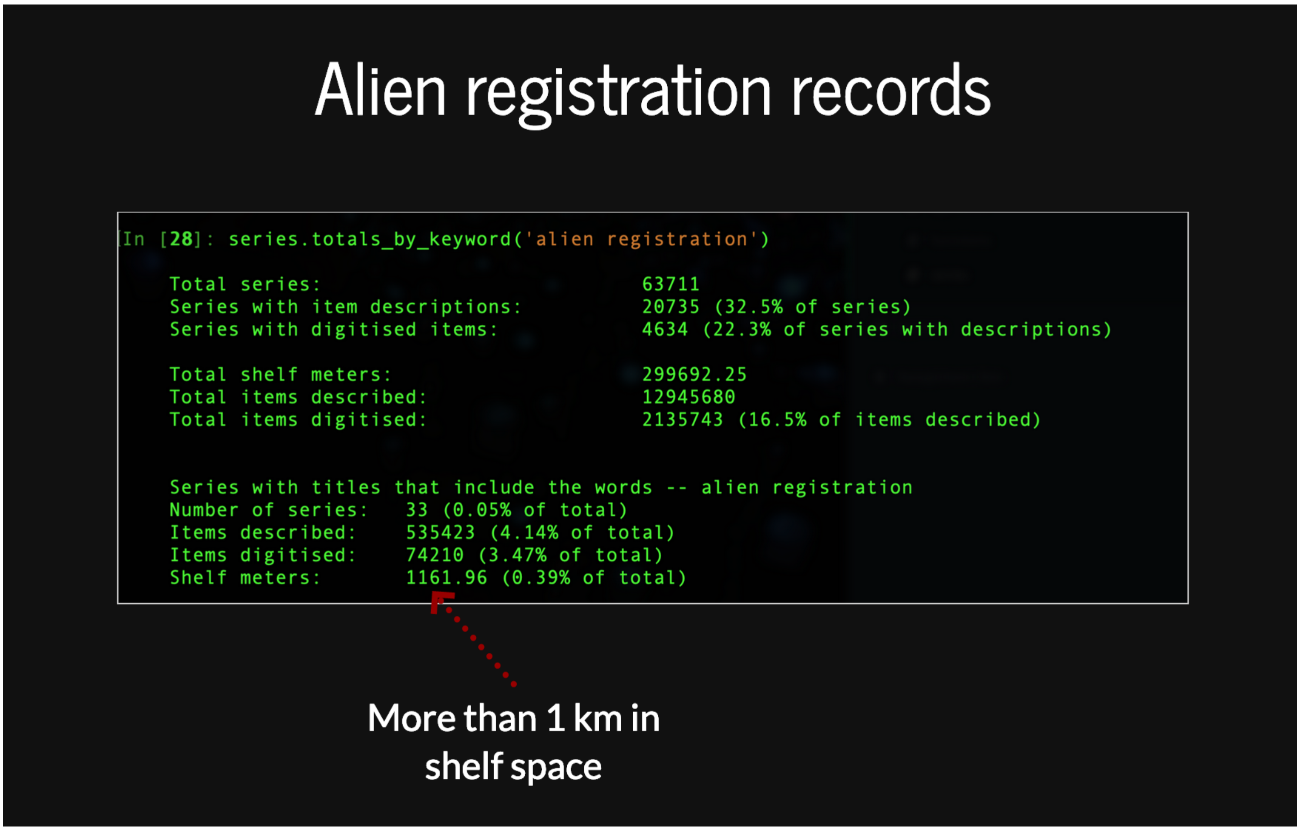


The words that define who can enter, and who belongs, also justify continued surveillance and control. Here's what we see if we look at the words that come *after* 'aliens' in newspapers. Systems of 'alien registration' were introduced during the two world wars and continued until 1971.



Many of the records created through this system are held by the National Archives of Australia. RecordSearch, the National Archives online database, has no API, but guess what — I have a few notebooks and screen scrapers that can help you get data out.

Having harvested most of the series-level data from RecordSearch I can now aggregate it by a keyword or phrase. So I can tell you that series with the phrase 'alien registration' in their title take up more than one kilometre of shelf space in the National Archives.



Alien registration records

```
In [28]: series.totals_by_keyword('alien registration')

    Total series:                           63711
    Series with item descriptions:          20735 (32.5% of series)
    Series with digitised items:            4634 (22.3% of series with descriptions)

    Total shelf meters:                     299692.25
    Total items described:                  12945680
    Total items digitised:                  2135743 (16.5% of items described)


    Series with titles that include the words -- alien registration
    Number of series:   33 (0.05% of total)
    Items described:    535423 (4.14% of total)
    Items digitised:    74210 (3.47% of total)
    Shelf meters:       1161.96 (0.39% of total)
```

More than 1 km in shelf space

But what if they don't have that term in their in title. Can we trust the language of our descriptive systems?

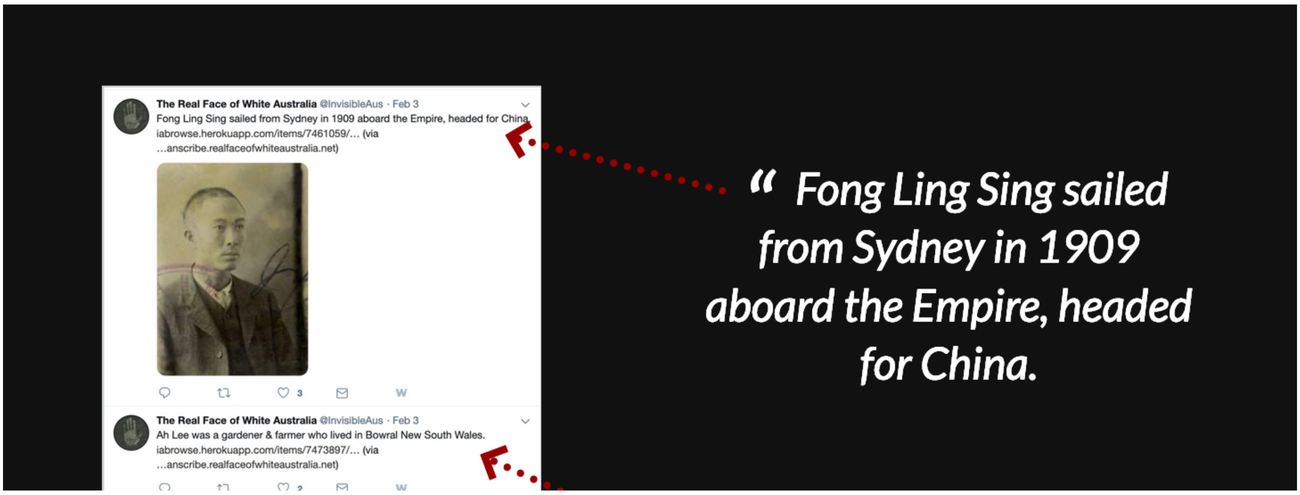
If you search in RecordSearch for series descriptions that include the term 'White Australia Policy' you get very few results. There are many, many records relating to the surveillance and control of non-white populations in the National Archives of Australia, but the descriptions rarely include that phrase. You have to know the language of the racist bureaucracy that created the system in order to find the records.

Our own descriptive systems, our own language, our own words can maintain the invisibility, the otherness, of people marginalised by past governments.

Using and sharing these sources, analysing their contents, and looking for patterns — these all help cut across the biases of our descriptive systems. While searching for 'White Australia Policy' in the National Archives is not very productive, I can harvest and package data about relevant series and share them in a public repository.

Or I can take data scraped from early naturalisation applications in the Archives and create a Twitter bot that just tweets their names and links to their files — many names, many cultures, these were the people of Australia.

The Real Face of White Australia Twitter bot uses data transcribed from files used in the administration of the White Australia Policy to create little stories about the people who lived under its restrictions. These were the people of Australia.



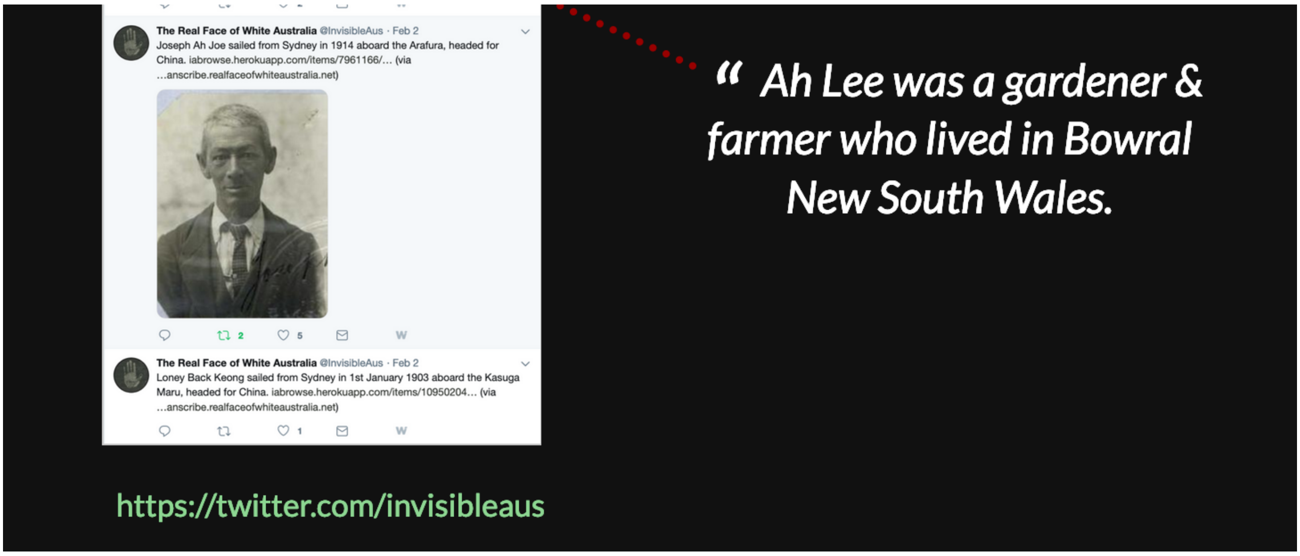
The Real Face of White Australia @InvisibleAus · Feb 3
Fong Ling Sing sailed from Sydney in 1909 aboard the Empire, headed for China. iabrowse.herokuapp.com/items/7461059/… (via …anscribe.realfaceofwhiteaustralia.net)

3

The Real Face of White Australia @InvisibleAus · Feb 3
Ah Lee was a gardener & farmer who lived in Bowral New South Wales. iabrowse.herokuapp.com/items/7473897/… (via …anscribe.realfaceofwhiteaustralia.net)

2

" Fong Ling Sing sailed from Sydney in 1909 aboard the Empire, headed for China.



The Real Face of White Australia @InvisibleAus · Feb 2
Joseph Ah Joe sailed from Sydney in 1914 aboard the Arafura, headed for China. iabrowse.herokuapp.com/items/7961166/... (via ...anscribe.realfaceofwhiteaustralia.net)

The Real Face of White Australia @InvisibleAus · Feb 2
Loney Back Keong sailed from Sydney in 1st January 1903 aboard the Kasuga Maru, headed for China. iabrowse.herokuapp.com/items/10950204... (via ...anscribe.realfaceofwhiteaustralia.net)

“ Ah Lee was a gardener & farmer who lived in Bowral New South Wales.

https://twitter.com/invisibleaus

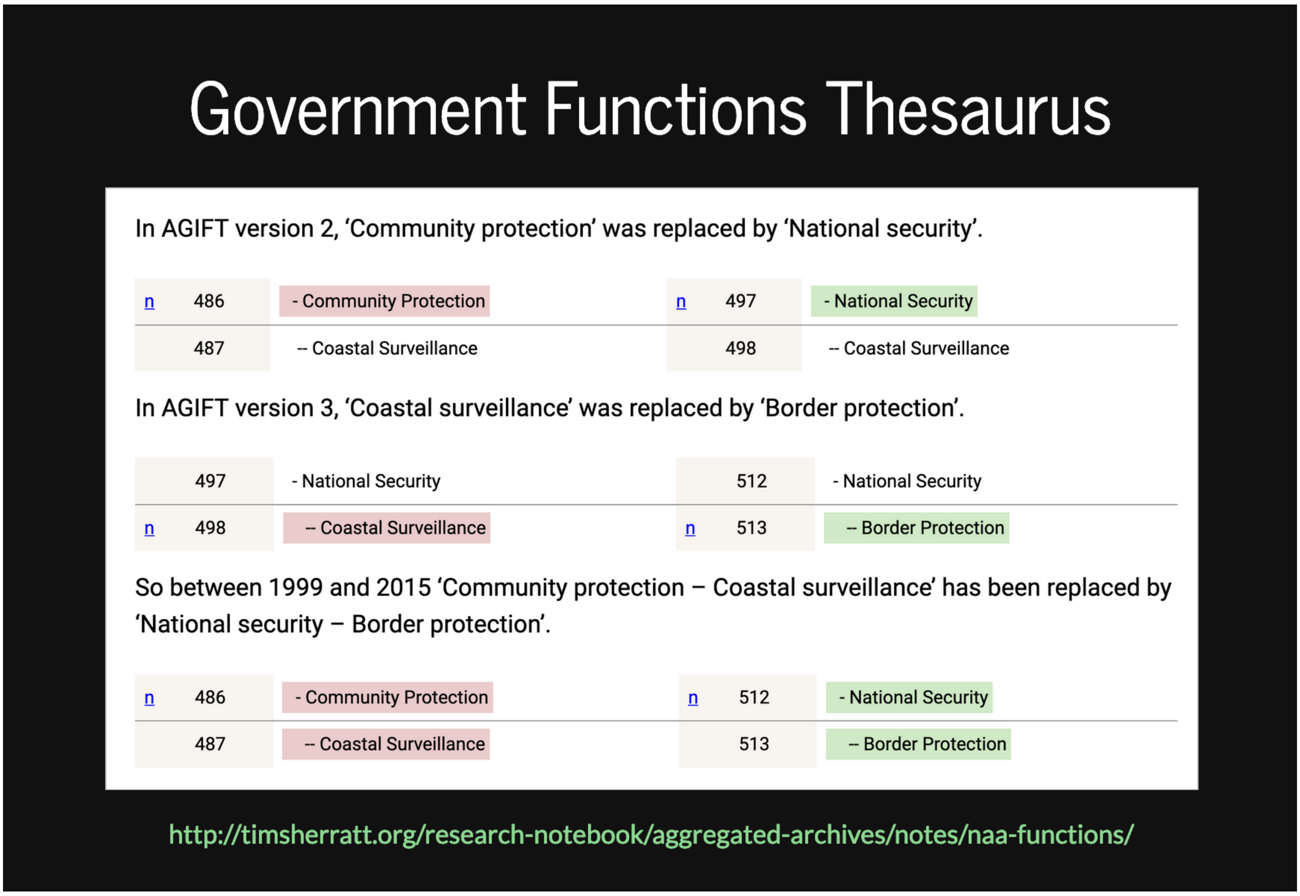
@InvisibleAus

In recent years, the Australian government invented a new category of ‘Illegal Maritime Arrivals’ to justify the brutal treatment of asylum seekers. Amongst its planned changes to the citizenship test is a much higher level of English language skills. Last year a senator from Queensland suggested a return to the White Australia Policy would not be a bad thing.

A couple of years ago the UNHCR launched the #wordsmatter campaign to remind people in well-off countries like Australia that asylum seekers were not ‘illegal’, and refugees were people forced to flee their homes due to danger or persecution.

Words matter — as people who work with text, who describe collections, who create and consume metadata, we know this. But still…

The National Archives of Australia maintains a hierarchical thesaurus of government functions — things like Defence, and Education. It’s intended to aid the discoverability of online resources, but there’s also a version used in RecordSearch itself to help describe government agencies. I wasted a lot of time comparing versions of this thesaurus, but one interesting thing I noticed was this.



Government Functions Thesaurus

In AGIFT version 2, ‘Community protection’ was replaced by ‘National security’.

n	486	- Community Protection	n	497	- National Security
	487	-- Coastal Surveillance		498	-- Coastal Surveillance

In AGIFT version 3, ‘Coastal surveillance’ was replaced by ‘Border protection’.

	497	- National Security		512	- National Security
n	498	-- Coastal Surveillance	n	513	-- Border Protection

So between 1999 and 2015 ‘Community protection – Coastal surveillance’ has been replaced by ‘National security – Border protection’.

n	486	- Community Protection	n	512	- National Security
	487	-- Coastal Surveillance		513	-- Border Protection

http://timsherratt.org/research-notebook/aggregated-archives/notes/naa-functions/

More details at http://timsherratt.org/research-notebook/aggregated-archives/notes/naa-functions/

Between version 1 and version 3 of this thesaurus, the function: ‘Community Protection — Coastal Surveillance’ was replaced by ‘National Security — Border Protection’. You could argue that this is in line with changes to the names of the relevant government agencies — instead of customs officers we now have Border

Force. But agencies change names all the time. Has the function itself actually changed?

Words matter. They help enforce boundaries and define who belongs. Whose borders are we actually protecting?

Share **Tweet** **LinkedIn** **Reddit**

Previous
← *2018 — the making and the talking*

Next
Trove: connecting us to the past →