4 million articles later...

Author : Tim Sherratt

Tagged as : frontpages, newspapers, Trove

Date : June 29, 2012


The Sydney Morning Herald.

GIGANTIC AIR OFFENSIVE AGAINST EUROPE

On 15 April 1944 the *Sydney Morning Herald* turned inside out. For more than a hundred years, the front page had been dominated by advertisements, but this changed suddenly in 1944 as the newspaper took on a completely new look. In place of the ads were the day's top stories, headlines and photographs — a 'front page' design familiar to modern readers.

The change was, the newspaper explained, partly a response to the demands of war. Advertising had been cut due to the rationing of newsprint and 'an urgent public demand in these critical days for more papers and more news'. But they were also looking forward to the problems of peace:

> It is essential… that we should not only provide the space, but also adopt the manner and methods of presentation which will spread knowledge of these problems yet more widely, and bring them home yet more deeply, among the people of this country.

But the *Sydney Morning Herald* wasn't breaking new ground. The design of front pages had been changing across the first half of the twentieth century as advertisements gradually gave way to news. This graph shows the average number of words per issue on the front pages of Australian newspapers devoted to advertising.

You can see a clear decline from about the turn of the century. News articles, on the other hand, were on the way up.

Not all the changes were as sudden as the *Sydney Morning Herald*'s. *The Barrier Miner* entered the First World War with the ads on top, but by war's end the position was reversed. In between was a period of transition as you can see from this graph which plots advertising against news.

If you dig a bit deeper, you find that the amount of advertising follows a regular pattern.

These peaks and troughs in June 1916 are a week apart — Saturday's front page was all advertising , but the next day brought a 'Special Sunday Issue' focused on the 'Latest War News'.

It's clear just from these two examples that there are stories behind these changes. There are subtleties and contingencies to be explored along with dramatic shifts.

And now you *can* explore them…

The Front Page

The Front Page is a database containing details of more than 4 million front page newspaper articles harvested from the National Library of Australia's Trove service.

Trove divides articles into a series of categories:

- articles (news)

- advertising
- detailed lists, results, guides
- family notices
- literature

I've simply gone through and added up the numbers of articles and the numbers of words in each category for each issue, and aggregated this across months, years and the full run of each newspaper.

These totals are presented as a series of linked tables and graphs. Just click on a point to zoom in, or use the navigation controls to go directly to the issue of your choice. It's pretty straightforward.

Why?

We're lucky to have rich resources like Trove, but if we're going to make best use of them we have to move beyond the search box to find new ways of exploring and contexualising their content. That's why I've developed tools like QueryPic, Headline Roulette and even The future of the past. Each lets you engage with the newspaper database in a different way.

But not all newspaper articles are created equal. I'd like to be able to aggregate and analyse the 'top' stories for each day, but to do this I need to know more about the structure of the newspapers themselves. I've already made a few attempts to find and extract editorials. This is useful because before the main news moved to the front page it was often directly after the editorials. But when did the news shift to the front page?

Now I can find out.

But why create a public web resource? Well, it's just what I do. I build and I share. It's what motivates me. It's how I understand things. It's where I find both my questions and my answers. Hey, I'm a digital humanist ok?

How?

Everything's up on GitHub, so you can follow along with my ugly coding. It was all a bit of an experiment, because I simply didn't know whether I *could* harvest and use 4 million articles. How long would it take? Would MySQL grind to a halt? Would my laptop blow up?

In my Harold White lecture I wondered whether what I was trying to do was really beyond the reach of 'an ordinary bloke and his laptop'. I suspect the day is rapidly coming where my work will be superceded by well-funded academic projects with access to supercomputers and a pool of bright young graduate students. But for now I'll just keep pushing the boundaries of what's possible over a dodgy home broadband connection.

Of course, this project was only possible because of the Trove API. My screen-scrapers of yore would have been impossibly slow and wasteful of bandwith. With the API I could simply construct a query and then loop through the 4 million articles in batches of a hundred. These were then fed into MySql via Django. I quickly worked out that I needed to keep my Django models simple. My clever relational model linking newspapers, issues, pages and articles was just too complex for this sort of operation. I flattened everything out to store all the metadata in a single 'article' model.

The harvesting operation took about 5 days. Once I had all the metadata I ran a couple of processes to do all the adding up and saved the results to a separate 'totals' table.

Then it was just a matter of building a front end. Using Django, Twitter Bootstrap and HighCharts made this amazingly easy. Really. Really truly.

What now?

I built this because I wanted to track changes in the design of front pages, but now I'm wondering what else I can find. The role of war in the examples above is intriguing. Are there other changes in our relationship to 'news' that these graphs might reveal?

I hope other people will wonder about this as well.

I have some ideas for future developments. For example, I'd like to add tagging to make it easy to construct timelines of significant changes. But first I just want to see if anybody's actually interested. If you have any ideas, suggestions or comments please let me know.

Ok, off you go — explore.