Hacking heritage: power and participation in digital cultural collections

Author : Tim Sherratt

Tagged as : hacking, Hansard, invisibleaustraliansrecordsearch

Date : July 15, 2016

Presented at the DigitalGLAM Symposium, University of Melbourne, 15 July 2016.
Play with the slides*.*

Who's used Commonwealth Hansard through the ParlInfo Database — particularly the early stuff, pre 1980?

Was it fun?

It's great that all of that content has been digitised, but the user interface is not very friendly — one of the main problems is that it's just really hard to read anything in context, to get a sense of proceedings in parliament as they unfolded.

However, you can bypass the interface and go straight to the data. There's an XML file — with the proceedings in a nicely-structured, machine-readable form — for each day.

To save the world much searching and clicking, I wrote a script to crawl through ParlInfo and download all of those XML files. There's about 4gb in total, and I've shared them all through my GitHub account. Feel free to download and explore!

I initially harvested the files because I thought they'd provide a useful example for use in teaching about text analysis tools and techniques. But then I reflected on that 'readability' question and wondered how much work would be involved in creating a version of Hansard that was focused on browsing and reading. The answer was about two days — over the space of a weekend I created Historic Hansard.

I wanted to mention this today because I think it's a good example of a 'hack'. Contrary to mainstream media coverage, most hackers don't want to steal your credit card details. Hackers, within the diverse community of coders, builders, programmers, and developers, are people who use code to solve problems, to get around obstacles — their solutions might not be the most elegant, but they work.

My bio reads 'historian and hacker'. I'm a historian who uses digital tools and techniques to get around obstacles. In particular, I use code to open digital cultural collections.

In the case of Hansard I have, for minimal cost, created something that allows you to explore the proceedings of parliament in context. It's made some historians pretty excited. Since that first weekend I've added an index to people and bills. I've also integrated tools for text analysis and annotation. Shortly I'll add the rest of the House of Reps and the Senate through until 1980. I was very pleased that the new PMs site from the Museum of Australian Democracy was able to harvest from Historic Hansard to boost their content.

So this is an example of what happens when governments or cultural heritage organisations provide openly-licensed data for people to play with — although I should note that the contents of the parliament website, including Hansard, have a 'no derivatives' licence which creates a lot of uncertainty around what you can actually do.

I've talked a lot in recent years about building stuff with cultural heritage data — particularly in relation to Trove. And I've created a lot of tools, experiments and interfaces — some significant, some silly, some downright creepy.

But while open data does enable mildly-obsessed hackers like me to build new, shiny things, it also enables us to ask different questions — to analyse what's not there, as well as what is.

In the case of Commonwealth Hansard what's not there is about two years' worth of Senate proceedings. I stumbled on this while harvesting the XML files and started doing some analysis. Between 1901 and 1980, 94 days of Senate proceedings, and 8 days of the House of Representatives are effectively invisible. PDFs exist, but the XML files are empty and the content is not searchable.

As you can see, most of the 'invisible' days come from the WWI period. So if you've been relying on the online version of Hansard to research WWI, you probably need to think about what you might have missed.

Needless to say, it's unlikely that anyone would have noticed these gaps using ParlInfo's web interface. It's yet another example of why you always be suspicious of search engines. But it's also an example of how hacking collections opens them to new forms of analysis and enables us to critique the very notion of access.

Spend sometime reading the strategic plans of cultural heritage organisations and you'll see that 'access' figures prominently — access is 'provided', 'given', and 'opened'. Online access is a 'deliverable' to be measured in hits and sessions.

The problem with this is it casts the public as consumers of access, rather than creators. Hacking heritage collections is one way demonstrating that access is not a one way flow of information — it's a struggle and it must always be. We should always be uncomfortable with the categories used to structure our past. We should always want more. If we run out of questions and criticisms then something's gone badly wrong.

Who's used the National Archives of Australia?

You've probably come across the process of 'access examination'. Records more than 20 years old are assessed against a list of criteria to see if they can be released to the public — the criteria are defined under the Archives Act and relate to things like national security and privacy.

Most records end up with an access status of 'open', but some end up 'closed' and are withheld from the public. Last year I started wondering what happens when we flip around the idea of 'access' and focus on the files we're not allowed to see?

The National Archives' online database, RecordSearch, provides basic metadata about 'closed' files, including the reasons why they've been withheld. But you can't easily search or analyse this data within RecordSearch itself. So I wrote another script that harvested details of all 14000 files with the access status of closed and created my own interface to them — Closed Access.

While my Hansard hacking was aimed at opening up new forms of use, Closed Access is about understanding the processes and practices that determine whether we're allowed to see a file. It pushes beyond legislative definitions in an attempt to reveal 'access' as a process that is historically contingent — there is nothing magical, mysterious, or dangerous about a 'closed' file.

By hacking collections we can also see them differently. Amongst its millions of records the National Archives holds the remnants of the White Australia Policy. About 5 years ago I wrote a script to harvest several thousands of these documents. I ran them through a facial detection script, cropped out the portraits, and created this seemingly endless wall of faces.

It was another weekend project, another hack. In this case it exposed the remnants of a racist bureaucratic system in a completely different way. Instead of databases, record series, and files, you could see the people inside.

Of course the danger in raving about the wonders of hacking is that we simply create new structures of power, new limits on participation based around coding ability. I don't believe in hacker elites liberating the masses from ignorance. So part of the struggle around access must also be the struggle to share — to share technology, code, examples, tools, data, knowledge, possibilities, and passions. To give people the opportunity to participate and not just consume.

You might think sharing is easy, but it's fucking hard work — both for institutions and individuals. And it's work that is rarely rewarded in terms of professional acknowledgement, advancement or funding.

But I'm inspired by open access activists around the world, by radical librarians setting up Tor exit relays, by digital humanists critically examining unequal access to technology, and exploring the possibilities of minimal computing. By all the people who freely share their work and ideas to make 'access' just that little bit easier for others.

I share because I think it's important, because it's part of the obligation of being a hacker and historian, academic and activist.

So I share things like:

- articles and examples on my blog
- code and data on GitHub
- current research notes and hacks in my research notebook
- documentation, activities, and tutorials in my digital heritage handbook
- web apps and sites all over the place

I'm also trying to develop a few DIY, low code examples that enable people to build things using Trove data and some freely available tools. I'm hoping that'll give them the confidence to explore further. A starting point perhaps…

Just last week I ran a free digital heritage workshop in the lead up to GovHack — the national, open

data competition. We're hosting a specially-themed 'Heritage Hack' node at the University of Canberra as part of the GovHack this year. Feel free to join us!

Not that any of this is particularly special or remarkable. It should just be what we do — as researchers, as cultural institutions, as citizens, as hackers. The main thing we need to share is, of course, our power.

So let's just just stop talking about access to cultural heritage as if it's a destination, and explore it as journey full of tensions, dangers and difficulties; but also full of joy, discovery, and meaning. Let's hack heritage to build critiques as well as websites, conversations as well as code.

Share this:

- Click to email this to a friend (Opens in new window)
- Click to print (Opens in new window)
- Click to share on Twitter (Opens in new window)
- Click to share on Facebook (Opens in new window)
- Click to share on Google+ (Opens in new window)
-