The practice of play

Author : Tim Sherratt

Tagged as : APIs, data, play, recordsearch, research, screenscrapingTrove

Date : February 25, 2017

Keynote presentation at the Deakin University Faculty of Arts and Education HDR Summer School*, Geelong, 24 February 2017.*

I'm a historian. But in the past decade the nature of my research has changed quite profoundly. Instead of heading off to the archives, taking lots of notes, and writing up a book or an article, I now make things. Generally these things are online, and open to the public. I make things for people to use, to explore, to play, and to ponder.

I started down this track before I realised there was a name for what I do – practice-led research. The things that I make even have their own acronym – they're NTROs, or Non Traditional Research Outputs.

But practice-led research is not just about making things. New knowledge is generated through cycles of creation and reflection. My aim in making is not to follow a blueprint, or check off a list of requirements, but to end up asking 'What is this thing?', 'What does it do?', '*How* does it do it?'.

In the past, I've tended to talk about my research practice as *playing* with data. I think there's an important argument to be made for the role of play in research, particularly when confronted with large cultural datasets. But 'play' doesn't quite capture what I do, nor does it look very convincing in a research proposal. So what do I really do?

Let's play a game.

Headline Roulette is a very simple game. Presented with the title of a digitised newspaper article drawn at random from Trove's collection of more than 200 million you are challenged to guess the year in which the article was published. Sounds easy, but you only get ten guesses. It's sort of like a cross between hangman and *The Price is Right*.

Despite its simplicity, I've known it to unleash the competitive instincts of a workshop full of historians. But for me, Headline Roulette is important because it provides an example of what becomes possible once we make cultural heritage collections available online. Our interactions are no longer limited to conventional modes of viewing or reading – we can play, and we can build.

I made the first version of Headline Roulette back in 2010. It was a game, but it was also an argument about access and possibilities.

Perhaps we should first take a step back. Who's used Trove?

Trove is a fundamental part of Australia's research infrastructure – and not just for those of us in the humanities or social sciences. Trove is a lot more than digitised newspapers, but access to more than 150 years worth of digitised newspapers has profoundly changed historical practice.

I say this not just because you have been spared the pain and suffering wrought by microfilm readers upon a generation of historians, but because the meaning of access itself has changed. Headline Roulette is just one simple and silly example of how once cultural heritage resources are in digital form we can use them differently. We can see them differently.

Imagine your search in Trove's newspapers zone returns 10,000 or 100,000 results. How do you make sense of that? How do you get an understanding of the whole, when all you see is page after page of search results?

QueryPic extracts data from Trove to visualise your search as a single chart – showing you the number of articles per year that match your query. You can even compare the occurrence of particular words or phrases.

But that's only the beginning, because once you think about web resources as data rather than just another type of publication you can aggregate and analyse – you can look for big, dramatic pictures as well as tiny, fragile fragments.

Trove Harvester is tool that delivers historical newspaper articles in bulk – thousands, even millions of articles saved to your computer for offline exploration.

What might you do with a million newspaper articles?

Research using digital resources like Trove is not constrained to the window of your web browser. You can ask new types of questions.

But back in 2010–11 when I created the first versions of Headline Roulette, QueryPic and the Trove Harvester there was no easy way of getting data out of Trove. The thing is, web pages are good for delivering data to human beings, but not so good for computers. Computers are actually pretty dumb, and you need to be quite explicit in packaging up data for them. Nowadays Trove has a thing called an API (an Application Programming Interface) which delivers data in a carefully structured format that even computers can understand. You can use APIs to harvest data, or to build new tools or interfaces. APIs are cool.

Without an API, the first versions of my tools had to turn human-readable web pages into computer-readable data – a process known as screen scraping. They were, therefore, not only useful or interesting applications in their own right, they were arguments about why things like APIs matter. Why web pages aren't enough. Why researchers need access to data.

These are arguments we're still making. Next week I'm heading to a workshop in California where we'll be discussing how libraries and other cultural institutions can deliver their data in ways that support new forms of research.

But we don't have to wait. By screen scraping web pages, by reverse engineering online databases, we can continue to develop the argument for access by extracting, sharing, and using data.

What could you do with 70gb of digitised surveillance files from the Australian Security Intelligence Organisation (ASIO)? I'f you'd like a copy I have them here on a USB drive.

Don't worry – we're not about to be raided by the security services. These are all files that have been carefully examined and released to the public through the National Archives of Australia. You can find them by searching the Archives' online database – RecordSearch.

Who's used RecordSearch? It's not the most friendly system, but the collection it documents, and

the metadata it provides, is rich and wondrous. I've spent a lot of time trying to get useful data out of RecordSearch – not just ASIO files, also records documenting the administration of the White Australia Policy, as well as higher-level data aimed at building my understanding of how the Archives, and its descriptive systems, actually work.

It is painful and frustrating work. But, I would argue, **it is research**. Terms like 'data mining' and 'text mining' fly around all the time, making it seem as if the the accumulation of data is a mechanical process – as if we're just digging it up. But the practice of screen scraping, or of liberating data from any cultural heritage source, is not simply extractive – it's iterative and interpretative. It's a process through which you begin to understand how the data is organised, what its limits and assumptions are, what its history is. What it means. We're not just taking things out, we're putting them back.

Frederick Gibbs and Trevor Owens argue that historical data need not be deployed solely as statistical evidence. 'It can also help', they suggest, 'with discovering and framing research questions' – questions, not answers; interpretation not calculation. Gibbs and Owens describe an 'iterative interaction with data as part of the hermeneutic process'.

For me, RecordSearch is like an archaeological site. Excavating data from it involves digging through layers of technology, institutional history, and descriptive practice to try and understand why we have what we have.

Those of you undertaking projects using the collections of the National Archives will almost certainly come across the process of 'access examination'. Under the Archives Act, government records more than twenty years old are expected to be opened to the public. However, the act also defines a number of exceptions to this rule – for example, records that endanger national security or infringe an individual's privacy can be completely, or partially, withheld from scrutiny. The process of assessing records against this set of exemptions is called 'access examination'.

The vast majority of records are opened without problem – they are, after all, more than 20 years old. But a significant number are not. While you can't use these records, RecordSearch does provide some information about them. So I decided to see what we couldn't see.

In January 2016 I fired up my screen scraper and harvested details of all the files in RecordSearch that have the access status of 'closed' – there were 14,370 of these files that had been through the process of access examination and withheld from public view. I then created my own interface that lets you explore this data from a variety of angles – such as the reasons why files were closed, when decisions were made about them, how old they are, and which government agencies created them.

It is perhaps the most frustrating search interface ever devised, given that you're not allowed to see any of the files you find.

Those of you currently planning research projects might be interested to know where most of these files come from. It's not defence or the intelligence agencies, but what is now the Department of Foreign Affairs and Trade (DFAT) – in January 2016, there were 1,747 closed files from just one DFAT series. But if you dig deeper you see that most of these files aren't withheld for one of the reasons defined by the Archives Act, they are described as 'closed pending advice'. The National Archives is still waiting to hear back from DFAT about them. Using my interface you can see that there were 54 files in this series where the Archives has been waiting for more than five years. So if you're embarking on a project using the National Archives, make sure you get your access examination requests in early. Just in case.

My aim in extracting and sharing this data is to better understand access examination itself as a historical process. It's work that enables us to ask different types of questions, but it also makes a change in the process itself. My interface is public, offering a critical commentary on the 'official'

system. As a result of my research, the Archives has made changes to the way it describes closed files. It's both research and intervention, history and hack.

'Hack' has a number of definitions, both positive and negative. Mark Olsen describes the 'hacker ethos' as:

> 'a way of feeling your way forward through trial and error, up to and perhaps beyond the limits of your expertise, in order to make something, perhaps even something new. It is provisional, sometimes ludic, and involves a willingness to transgress boundaries, to practice where you don't belong… Whether eloquent or a kludge, a hack gets things done.'[1]

Olsen explores what hacking means in the context of the humanities, arguing not only that hacking has a legitimate place in humanities practice, but that the humanities itself needs to be hacked to foster the development of new skills and literacies.

At this point you're probably thinking, 'But I don't do any of this wacky digital stuff, what has this got to do with me?'

Who's heard of filter bubbles, or search personalisation? Who's read one of the many reports recently about the way computer algorithms are shaping our online experience? Olsen argues for a humanities practice that equips us to wrestle with complex techno-social systems.

And we're not just talking about Google.

Last year Matthew Reidsma published an analysis of algorithmic bias in library discovery systems. He hacked a common commercial library product to show some of the biases underlying its recommendations system. The interfaces we use to access information are never neutral. The databases we search are products of selection and exclusion. Hacking enables us to interact with these systems as critics, and not just consumers.

Using the Trove API you can create a chart showing the number of digitised newspaper articles available per year from 1803 onwards. If you do this, you'll notice two significant features. First, there is a dramatic drop-off in the number of articles after 1954. This is the 'copyright cliff of death'. Few things are certain in our overly-complex copyright system, but 1954 provides a practical cut-off point. History stops in 1954.

You'll also notice a substantial peak in the number of articles around 1914. Why might this be? Did something significant happen in 1914?

In fact, it's all about money. In the lead up to the centenary of WWI it was decided to focus limited digitisation resources on newspapers from the WWI period. It was a perfectly reasonable decision, but the consequences are effectively invisible to any user of the web interface. You don't know what you're searching.

The power of Google encourages us to put a lot of faith in search interfaces. We trust that they will *just work*. And if we can't find what we're looking for, we often assume that it doesn't exist.

Hansard, the recorded proceedings of the Australian parliament from 1901 can be searched using the ParlInfo database on the Australian Parliament House website. Perhaps you've used it – it's a wonderfully rich resource. Powering the search results are a series of well-structured XML files, one for each sitting day, that identify individual debates and speeches.

Last year I reverse-engineered ParlInfo and harvested all those XML files. I thought they'd provide a great dataset for exploring changes in political speech, and so I created a repository containing all the files for the House of Representatives and the Senate from 1901 to 1980. Feel free to download and play.

But in the process of harvesting the files I noticed that some of the XML files were empty. After a bit more analysis I realised that about 100 sitting days were missing – they didn't show up in search results on ParlInfo.

The 'missing' days were concentrated in the Senate between 1910 and 1920. So anyone relying on ParlInfo to research the WWI period would have missed significant amounts of content. This 'black hole' was effectively invisible to any user of the web interface. It was only though hacking that its shape and extent was revealed.

Fortunately staff at the Parliamentary Library have investigated and fixed the problem. But it's a good example of why we should, as researchers, start from the assumption that search interfaces lie. Processes of selection and description shape the 'reality' of online collections. We then explore them through complex technological systems that appear comprehensive, even when they are not. You can't find what's not there. Online collections hide as much as they reveal.

Of course this is true of all historical sources. We are trained to analyse both context and content, to make judgements about authenticity and accuracy. These same skills need to applied to digital resources, to data. Indeed, Gibbs and Owen argue that 'historians must treat data as text, which needs to be approached from multiple points of view and as openly as possible'. But how do we find multiple points of view when interfaces construct our experiences and limit our perspectives. How do we open data to new possibilities? How do we see data differently?

No doubt you've been encouraged to find a way of expressing your research questions succinctly, in a way that communicates with a non-specialist audience – yes, I mean the dreaded elevator pitch. You're not the only one.

I've landed back in academia after a number of years working in cultural heritage institutions, and pursuing my own research interests with the support of the international digital humanities community.

Believe me when I say, Twitter changed my life. There I was, hacking away on cultural heritage data without any real assistance or encouragement, when I discovered, via Twitter, that there were people out there like me. Many of these people are now my friends, and I've been lucky to travel around the world to meet and work with them.

But coming back to academia I've found that my collection of projects, tools, experiments, and obsessions was not quite enough – my research needs a 'narrative'.

So, like you, I've had to think about why I do what I do. What motivates my research? What matters?

For me it comes back to the nature of this thing we call 'access'. Cultural heritage organisations talk about 'access' all the time, particularly in relation to online collections. But what does it actually mean? I want to overturn our assumptions about access – exploring it not as a process of opening things up, but as a system of controls and limits. It's not a state of being, it's a struggle for meaning and power.

My methodology, and I *think* I can call it that, is the multiplication of contexts. Context is, of

course, critical to cultural heritage collections – it enables us to locate them within history and culture, to analyse their authenticity, to mobilise their value as evidence. But the descriptive systems we use to manage and explore collections represent only a privileged subset of possible contexts.

Now I'm still figuring this out, but I think what my work does is that it removes collections from these highly-controlled systems and lets them loose in a variety of new contexts. This allows unexpected features, or new uses, to emerge – we see them differently, and in that moment, the nature of access shifts, however slightly. It's those moments I'm trying to catch and observe.

If you've ever tried to use Hansard through the ParlInfo database you'll realise that it's just really difficult to read. You're presented with a series of nested fragments, so it's hard to get a sense of the context and flow of the day's proceedings. Having downloaded all those XML files, I thought I'd have a go at presenting Hansard in a form that privileged reading over search.

So I created Historic Hansard – dedicated to lovers of political speech. It does nothing very fancy, but I think it does it pretty well.

In the end, however, Hansard is still just text. What's lost in the documentation process is the performance – the theatre of parliament. But not completely. As well as formal speeches, many interjections have been recorded and preserved.

A few weeks ago I extracted all those interjections from 1901 to 1980, about a million of them, and saved them to a new database. As I fiddled with different presentation methods, I started to see them as something akin to tweets – quick, pithy, and pointed. What would happen, I wondered, if we reimagined interjections from a century ago in an age of social media.

Like many of my projects, this *whatever it is* took me a couple of days to build. No research grants were harmed in its creation, no committees were needlessly formed. This is not because I'm a whizz-bang coder – I'm certainly not. It has to do with the nature of this work – it's rapid, experimental, and sometimes even ephemeral. I don't design websites, I make interventions – things that are not only of the world, but in the world. They *do* something.

Stephen Ramsay explores the hermeneutical possibilities of screwing around with technology and texts. The 'screwmeneutical imperative' he suggests is based on the fact that:

> 'a writerly, anarchic text… is more useful than the readerly, institutional text. Useful and practical not in spite of its anarchic nature, but as a natural consequence of the speed and scale that inhere in all anarchic systems.

Digital technologies give us the opportunity to play with scale and speed. We can manipulate millions of newspaper articles, and we can build a new version of Hansard in a weekend. But this shift also applies to the *way* we communicate. Instead of waiting months or years for an article to appear in print, we can post it on a blog, or in a digital repository. It is fundamental to the work that I do that it is shared, it is public by default – not just the results, but the code, the data, the process, and yes the licensing. Access is not just what we take, it's what we do.

The multiplication of contexts has some interesting precedents as a research methodology. In the literary world the Oulipo movement sought to play with the contraints of composition. Lisa Samuels and Jerome McGann suggested that the deliberate misreading of a text, what they termed

'deformance', could yield critical insights. More recently, Mark Sample has argued for a 'deformed humanities' where we learn about things by breaking them.

In history we have the counterfactual – a creative reimagining of a past that never was, aimed at revealing perspectives and possibilities too quickly closed and forgotten. As Sean Scalmer argues, 'counterfactuals are fun':

> 'Conventions can be disregarded, or even mocked. Worlds might be remade, the tyrannical overthrown, and the humble elevated. New orders can be imagined.'[2]

But counterfactuals are not fiction. They work best when they sail close to an accepted version of the past; when they play with the constraints of documentary evidence rather than just ignore them. Just because an approach is playful, it doesn't mean that there are no rules. As Ian Bogost has recently argued, the fun of play is 'not doing what we want, but doing what we can with what is given'.[3] Play is an investigation of limits.

While some of the ASIO files held by the National Archives are closed to the public, most are 'open with exception'. This means that sensitive parts of the files have been removed. Whole pages can be withheld, or sections of text blacked out – a process known as redaction.

A redaction is, by definition, an absence of information, and yet the frequency, density, and placement of redactions across a large collection of documents could conceivably tell us something interesting. So last year I wrote a kludgy computer vision script that found and extracted redactions from digitised ASIO files. I now have a collection of 250,000 redactions which I've shared on Figshare – grab a copy now!

I'm continuing to explore the possibilities of these redactions as data points. But there was also something visually interesting about the redactions, particularly when they were assembled on masse.

Here you can browse all 250,000 redactions. But that's not all, you can also use them as entry points to the documents they were intended to obscure.

Contexts here have been reversed, the files have been turned inside out – the limits remain, indeed the scale of redaction is emphasised, and yet within these limits, perhaps even because of these limits, we can experience the files quite differently. We are no longer simply the subjects of state surveillance, we can reverse the gaze, inspect the process, and ask new questions.

The manipulation of contexts is not mere invention. The limits of access offer both meaning and rules. We have skin in this game, its outcomes matter, what is at stake is our ability to see, and be seen, within the cultural record. Access changes who we can imagine ourselves to be.

In the first half of the twentieth century, if you were deemed not to be 'white' and wanted to travel overseas from your home here in Australia, you had to carry special documents. Without them, you'd probably be stopped from returning – from coming home. This was 'extreme vetting' White Australia style.

Many thousands of these documents are now held by the National Archives of Australia. In 2011, I used my screen scraper to harvest about 12,000 images like this from RecordSearch. I then ran them through a facial detection script and created The Real Face of White Australia.

There are about 7,000 faces in this seemingly endless scrolling wall. And that's from just a small

sample of the White Australia records. It's powerful, compelling and discomfiting. But the power comes not from any technical magic, but from the faces themselves – from what we feel when meet their gaze. Once again the records have been turned inside out – instead of seeing files, metadata, or a list of search results, we see the people inside.

Play can be serious. It can make you feel things you don't expect. It can challenge your certainties and take you to the limits of what you know.

That sounds a lot like research to me.

1. M. J. Olson, 'Hacking the humanities: Twenty-first-Century literacies and the "becoming other"of the humanities', in Eleonora Belfiore and Anna Upchurch (eds), *Humanities in the Twenty-first Century: Beyond utility and markets*, Palgrave Macmillan, 2013, pp. 237–250. [↩]
2. Sean Scalmer, 'Introduction', in Stuart Macintyre and Sean Scalmer (eds), *What if? Australian history as it might have been*, Melbourne University Press, Melbourne, 2006, pp. 1–11. [↩]
3. Ian Bogost, *Play Anything: The Pleasure of Limits, the Uses of Boredom, and the Secret of Games*,Basic Books, New York, 2016, p. 236. [↩]

Share this: