A life reduced to data

Author : Tim Sherratt

Tagged as : invisibleaustraliansLinked Open Data

Date : August 25, 2016

Keynote presentation to the Migrant (Re)Collections *workshop, Leiden, 2016.*

In 1861, the census for the colony of New South Wales (as it was back then) recorded just one Chinese woman living in Balmain in Sydney. The historian Eric Rolls, writing in 1992, commented that this 'lone woman is exceptional and inexplicable'.

Inexplicable? My partner and collaborator Kate Bagnall is a historian of Chinese Australia and she recently investigated this case again, making use of digitised resources that were not available in the 1990s.

Her starting point was one tiny fragment in a digitised newspaper article on Trove. A report from the Water Police Court published in 1863 notes 'the case of Ah Happ, a Chinese woman, who claimed the sum of £8 9s 6d. wages for her services as nurse in the employ of Cyril Cecil, of Snail's Bay, Balmain'. The case was dismissed, but this brief, tantalising reference gave Kate enough information to trace the life of Ah Happ over the next 20 years or so, until she disappears from the records again. Kate now believes Ah Happ was the first Chinese mother in NSW. But was she the woman in the 1861 census?

The census has been big news in Australia recently, and not for the right reasons. I think the correct technical term for the handling of the 2016 census is *omnishambles* – there have been multiple failures both in communication and technology.

It all started to go wrong when the Australian Bureau of Statistics quietly announced that it would be keeping everybody's names for longer than usual and using them to generate identifiers that would link the census with other government datasets. This might not seem so controversial on its own, but of course context is crucial.

Over the last few years there have been multiple reports of the misuse of personal information by a variety of government agencies. Nonetheless the amount of information being gathered has increased. In 2015, for example, new laws were passed for the retention of metadata documenting the communications of all Australians. The census, previously a trusted tool for government planning, suddenly seemed a further creeping, encroachment on individual privacy.

As a historian with an interest in the politics of surveillance I had mixed feelings about it all. Concerns about data matching are well justified, but the census constitutes a vital historical resource – often documenting lives that are barely glimpsed through other sources. The controversy also overshadowed the fact that since 2001, a growing number of Australians had agreed that their name-identified census data could be preserved by the National Archives of Australia for release to researchers in 99 years. In 2011, more than 60% of Australians willinging added their details to the so-called Census Time Capsule. How much of that trust will have now been lost?

We're here to talk about questions of identity; to find better ways of matching records about people across historical datasets. So I think it's important to think about the how these datasets came to

be created. We are implicated in debates such as those that surrounded the recent Australian census. In some cases we are the beneficiaries of systems created for the surveillance and control of suspect populations. Time changes, but does not dispel, questions about our responsibilities to those we seek to *identify*.

Back in 2010 I wrote a blog post entitled 'I link therefore I am'. The National Library of Australia had recently established a service called People Australia which brought together a range of biographical sources, disambiguated the names of people and organisations, and minted persistent identifiers for each new aggregated identity. The service still exists as part of Trove. People Australia was also, I think, the first of the Library's online services to offer a public API.

This coupled with the development of Linked Open Data made me pretty excited. People Australia presented new opportunities to link resources across collections, but I was particularly interested in the possibilities for ordinary web users. After all, RDFa meant anyone with a web page could make Linked Data. There was also some activity at the time around machine tags – a sort of semantically-enriched tagging. Thanks to the vision of Aaron Straup Cope, machine tags were incorporated into Flickr in about 2009. They're still there – just…

To try and bring some of these threads together I created a simple web service that took a name and returned a snippet of nicely formatted, RDFa enriched, HTML that you could drop into your blog post or web page. A bookmarklet made the markup process relatively seamless. Alternatively, you could ask for machine tags that could be cut and pasted into services like Flickr. Instant Linked Open Data for the masses.

I gave a talk about all this to a group of librarians and challenged them to use my identity finder thing (I've never been good at naming things) to add machine tags citing People Australia identifiers to photos on Flickr – to unambiguously *identify* either the subject or creators of the photos. To encourage them further I created the Flickr Machine Tag Challenge, an interface that enables you to explore photos tagged with National Library identifiers. This was, I think, a very early example of crowdsourcing Linked Data.

With all this swirling around it was perhaps inevitable that questions of identity would figure prominently when Kate and I started to think about what to do with the large quantities of records held by the National Archives of Australia documenting the operations of the White Australia Policy.

For the non-Australians here – yes, when the Australian colonies came together to form a nation, it was generally agreed that the nation would be white, and that this would be achieved through the control of immigration. Of course the substantial Indigenous populations were conveniently ignored in all this. The Immigration Restriction Act was passed by the first Australian parliament in 1901.

But what about the thousands of non-white people already resident in Australia – Chinese, Japanese, Malay, Syrian and more. They were allowed to stay, but their movements in and out of Australia were monitored – they had to carry special papers that would exempt them from exclusion. Many thousands of these certificates are preserved in the Archives, documenting in ironic detail the lives of people who weren't supposed to be part of White Australia.

In 2010, Kate and I launched Invisible Australians – a project without any sort of funding or institutional support – aimed at drawing attention to these records. You might have seen one of our experiments – The Real Face of White Australia. It's a simple scrolling wall of more than 7000 faces extracted from the Archives using a facial detection script. It was a weekend hack that has been cited around the world. But the power, of course, is in the faces themselves – they confront us with the reality of Australia's racist past.

However, one of the main aims of Invisible Australians was to give names to those faces – to extract data from the archives that would enable us to link these tiny biographical fragments and follow

people through time. I'm about to have another look at this based on recent developments in crowdsourcing software – something like the Zooniverse's Measuring the ANZACs project would do a lot of what we wanted. But the project won't be the same as we imagined it back then. I've grown increasing uncomfortable with what it means to *identify* people.

In November last year, Mark Matienzo, the Director of Technology at the Digital Public Library of America, gave a paper in which he raised important questions about Linked Open Data and the 'power to name'. Like Mark I think we have an obligation to consider the contexts in which we create, recover, or aggregate identities. There is power in the process and we need to understand where it comes from and the violence it can do.

The question of identity was critical to the operations of the White Australia Policy. You might think that the certificates carried by non-white residents were nothing more than identity papers – an early form of passport. But the point is, only non-white Australians had to prove who they were. Moreover, the technologies used to determine identity – portrait photographs and handprints – were strongly associated with the management of criminal populations. Indeed, in 1911 one Chinese businessman objected to being treated 'just like a criminal'. The process of identification helped justify the racist underpinnings of the system – the management of this suspect group required special measures.

The taint of suspicion followed non-white residents through their daily lives. The Immigration Restriction Act created the category of 'Prohibited Immigrant' to describe those who were present in Australia illegally. Kate tells the story of one unfortunate cook in Melbourne who was arrested by Customs Officers and accused of being a prohibited immigrant, mostly because he seemed a bit too Chinese. He was forced to prove who he was and how and when he had come to Australia. This was something of a challenge as he always believed he was born in Australia, and had grown up in Chinatown after being orphaned at a young age. This story seemed all too relevant when Australia's immigration officials, now called 'Border Force', announced last year they would patrol the streets of Melbourne in a crackdown on visa fraud. Instead of 'prohibited immigrants' we now seek to identify 'illegal maritime arrivals'. And in the US, Donald Trump wants to introduce 'extreme vetting'.

I can't now look at our wall of faces without wondering about the uses of facial detection. There are easily available web APIs that will not only tell you if an image contains a face, but whether the face is smiling, its gender, and its race. Both Google and Facebook have claimed frightening levels of accuracy with their facial recognition technologies – not just finding faces, but matching them against a set of known identities. In Australia, a number of image databases are to be linked to create a new facial recognition service called – The Capability. Our faces are increasingly not our own – they are public signifiers to be captured by systems of identity management and surveillance.

This is the context in which we undertake our explorations of identity, in which we exercise our power to aggregate, and to name. We can of course turn these systems on themselves, in the way that the residents of East Germany claimed the Stasi archives as their own. There are a number of examples where archives of oppression have been reclaimed in the struggle for justice. But we have to make that decision and engage accordingly. There is no neutral position.

For me this means finding better ways of representing the uncertainties of identity. Technologies of surveillance construct identity as an aggregation of data points – matches, crossreferences, and hits. Linked Data tends to work the same way, mapping the points of connection across multiple datasets. We know the points do not make the person, but we use them to create a shell identity, and therein lies the challenge. How do we fill that shell with the complexities and contingencies of life, without losing Linked Data's ability to make meaningful and reusable connections.

As I mentioned, my early experiments with Linked Data were aimed at building simple tools that would give creators of content the power to enrich their work with structured data. Nowadays we

have tools like Pund.it and Hypothes.is that allow us to build layers of annotation and enrichment on top of existing web resources. We also have platforms like Scalar and the forthcoming Omeka-S that give us the ability to define relationships between resources within the context of interpretation. These sorts of tools help us bridge the gap between the data we collect and the stories we can tell.

But it should be easier. Over the years I've made a few attempts to combine historical narrative with Linked Open Data – all of them buggy and incomplete. But it's a project I keep coming back to for a number of reasons.

Firstly, historians create Linked Data all the time, they just don't realise it. In the process of their research they build complex entity-relationship models, linking people, places, events and resources. But when it comes time for 'writing up', the data gets squeezed out to fit with the conventions of linear narrative and print publication. We need new publishing paradigms that maintain the relationship between narrative and data and expose full richness of historical practice.

Second, one of the things that has always attracted me to Linked Open Data is the idea that anyone can create it, anywhere. Embed some RDFa, reuse some identifiers, and hey presto – you're publishing data about the world that can be aggregated and explored. At least in theory. In practice, developments in Linked Open Data seem to have been centralised around particular tools and institutions, or geared towards search engine optimisation. And yet, on the other hand, we continue to create specialised crowdsourcing platforms to foster public engagement with our cultural collections. Why not get better at sharing identifiers for collection items and support the development of simple Linked Open Data tools that can be used wherever content is created. Every blog post could become a collection portal.

And finally, because as much as I enjoy playing with data, stories are what really matters. Stories convey meaning and emotion in ways data cannot. They give us room to explore nuance and uncertainty. They make us human. I want to find better ways of enriching data with stories, and vice versa.

In 1908, James Minahan was declared a 'prohibited immigrant' and arrested. James had been born in Australia to a Chinese father and Irish-Australian mother, but when he was just five years old his father took him to live in China. When he returned 26 years later James spoke no English. How could somone born in Australia be a 'prohibited immigrant'? The authorities argued that James's connection to the country of his birth had been lost – culturally and racially, he could not be considered 'Australian'. But the case was hardly clear cut and eventually ended up in the country's highest court.

Kate has been researching the story of James Minahan for a number of years and has assembled a complex story of people, place, and law, drawn from the holdings of many cultural institutions. This time we're not going to let the data be squeezed out. Currently a draft of the story sits in a demonstration system I built using JSON-LD and AngularJS.

It works well enough, but I've decided to throw out most of the code and start again. Why? I feel that I've been focusing so much on the interface that the simplicity of the system has been compromised. I've been playing by Angular's rules and that makes me very uncomfortable. I've also been inspired by the work that's been going on in DH around the idea of minimal computing – creating tools that are both simple and sustainable; that demand little technical infrastructure, but build capacity for innovative digital research.

Ed., developed by Alex Gil and his team is a wonderful example – beautiful both in its aims and execution. Ed. is a theme for Jekyll, the static site generator, that makes it easy to publish scholarly editions of digital texts. I've decided to build a set of plugins and examples that will make it possible to add a Linked Open Data layer to a text in Ed. The result will just be HTML – easy to

publish, easy to preserve. Once I'm happy with the basics, I'll think again about adding some Javascript trickery to the interface. At it's simplest the data can be stored in a hand-edited YAML file. Not a triple-store in sight. There's something strangely liberating about working with Jekyll.

There have been a lot of mentions of 'sustainability' in the past couple of days. For me, sustainability isn't just about funding, or institutional support, or governance structures – it's about building things that can be hacked, reused, shared, and fixed.

Perhaps it's through systems such as this we can encourage and support the small-scale production of Linked Open Data, based not on machine learning or entity extraction, but on detailed research and individual expertise. The James Minahan story will not only explore the complexities of identity and belonging, it will map relationships between people, trace journeys through space, and provide a specialised subject guide to Australia's cultural heritage collections. At least that's the plan…

Perhaps we can find new ways of bringing together the microstories, that Marijke and others have mentioned, with the big pictures drawn from heritage data – of navigating changes in scale without losing sight of what matters.

And perhaps this mix of story and data can help us deal with the politics of identity more effectively – to share the power of naming, to provide space for uncertainty, to undermine the authority of those who seek to reduce us to a collection of data points.

I suppose you're wondering about Ah Happ. Was she the woman in the 1861 census? Unfortunately the dates don't quite match up, so it's hard to be certain. But given what Kate now knows of Ah Happ's history, the presence of the Chinese woman in Sydney in 1861 is hardly inexplicable. She was probably a domestic servant or a nursemaid. A data point left unnamed, but a person not unknown. Questions of identity are rarely that simple.

Share this: