Every story has a beginning

Author : Tim Sherratt

Tagged as : archives, history, history wall, invisibleaustralians, linked data, Mapping our Anzacs

Date : October 4, 2011

Entering the web of data

[view the presentation…] [view the triples…]

Keynote delivered at the annual conference of the Australia and New Zealand Society of Indexers, 14 September 2011.

This is me.

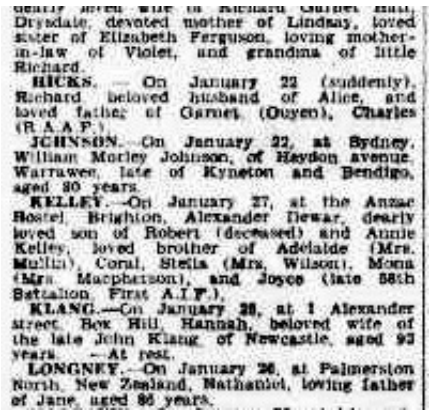
Today, Wednesday, 14 September 2011, I'm honoured to be able to join you here in the luxurious surrounds of the Brighton Savoy Hotel for the 'Indexing See Change' conference. This is an event, a moment in history; we can pinpoint ourselves, this gathering, both in time and in space.

If we do that, if we move outside the moment and position ourselves on a timeline or a map, interesting things start to happen. Connections emerge.

Here we are at number 150, The Esplanade, in Brighton. A bit over a kilometre away is the stately villa, Kamesburgh. For many years Kamesburgh was also known as the Anzac Hostel — a refuge for permanently-incapacitated World War One veterans.

The Anzac Hostel opened on 5 July 1919. Here it is draped in its patriotic finery, from the collections of the Australian War Memorial. According to the caption, the Anzac Hostel was 'a home, not an institute'.

Also amongst the War Memorial's holdings is a wheeled bed that was used at the hostel. This particular bed was apparently occupied by one man, Albert Ward, for forty-three years.



Drysdale, devoted mother of Lindsay, loved sister of Elizabeth Ferguson, loving mother-in-law of Violet, and grandma of little Richard.

HICKS. — On January 23 (suddenly), Richard, beloved husband of Alice, and loved father of Garnet (Queen), Charles (R.A.A.F.).

JOHNSON.—On January 22, at Sydney, William Morley Johnson, of Haydon avenue, Warrawee, late of Kyneton and Bendigo, aged 80 years.

KELLEY.—On January 27, at the Anzac Hostel, Brighton, Alexander Dewar, dearly loved son of Robert (deceased) and Annie Kelley, loved brother of Adelaide (Mrs. Mullin), Coral, Stella (Mrs. Wilson), Mona (Mrs. Macpherson), and Joyce (late 58th Battalion, First A.I.F.).

KLANG.—On January 28, at 1 Alexander street, Bex Hill, Hannah, beloved wife of the late John Klang, of Newcastle, aged 93 years. —At rest.

LONGNEY.—On January 26, at Palmerston North, New Zealand, Nathaniel, loving father of Jane, aged 86 years.

Death notice for Alexander Kelley. Argus, 29 January 1944.

It was probably in a bed just like this that Alexander Dewar Kelley passed away on 27 January

1944. Alexander Kelley was cremated, and his remains interred amongst the roses at what is now called the Springvale Botanical Cemetery. Not far from my own grandparents.

Alexander Kelley spent close to half his life in the Anzac Hostel. Like many young men, he bravely answered his nation's call to arms, but returned from war much changed. We can follow Alex's war through his service record, easily-accessible through the website 'Mapping Our Anzacs'.

Alex was a coach painter who enlisted in the AIF in January 1916. Within a year he was in France. In May 1917 he suffered a gunshot wound to the head, but was able to rejoin his unit in August. Less than a month later though, he was wounded again, this time more severely. For Alex the war was over, and he was shipped back to Australia in May 1918.

'Mapping Our Anzacs' includes a scrapbook feature through which visitors to the site can attach notes or photographs to a service record. Amongst the the many thousands of postings is a fragment from a diary, found tucked inside the bible of Alexander Kelley's mother. The diary entry reads simply: 'Alex arrived from Front. Wet day. Saw him at "Caulfield".'

Alex had survived and had returned to his family. This was a day to remember. But there was sadness too, for Alex was not the same young man who had left for the battlefields of Europe. In the diary fragment, 'Caulfield' is enclosed in inverted commas, indicating perhaps that the reunion took place, not in the suburb, but in the Caulfield rehabilitation hospital. Alexander Kelley was wounded in the face, hands and legs. He was left blind in both eyes and his right leg was amputated. He would live the remainder of his life a little over a kilometre away from here at the Anzac Hostel.

This is just one story. There are over 375,000 World War One service records held by the National Archives of Australia. How can we hope to understand a number like that? How can we hope to imagine the war's impact on families, on communities?

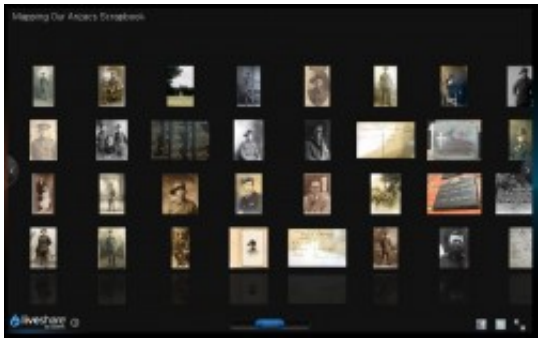
'Mapping Our Anzacs' uses familiar Google maps to display the places of birth and enlistment recorded in many of those service records. But technical limitations make it impossible to display all the places at once. You can, however, take the same data and open it in Google Earth. If you then zoom in on Victoria, you see something like this.

Mapping Our Anzacs data viewed in Google Earth.

Each marker represents a place where a service person was born or enlisted. It's impossible to read, of course, but that's the point. There is so little blank space. As you zoom further, more markers appear, more place names resolve. It's simple, but it's powerful. They came from everywhere. From the smallest village to the biggest city; nowhere was untouched.

The 'Mapping Our Anzacs' scrapbook offers another perspective. It's possible to extract the images

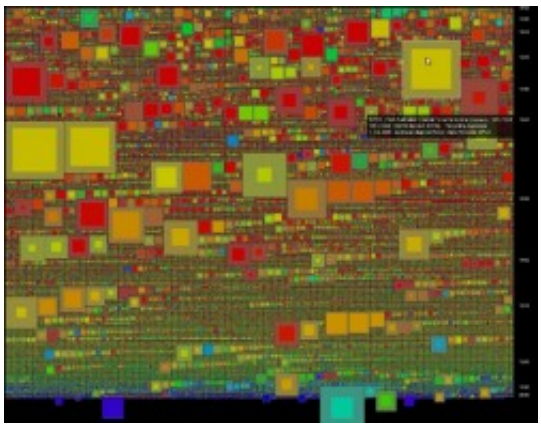
posted to the scrapbook and present them on a 3D wall. Amidst an assortment of memorabilia, there are faces. Not places, or records — this is a wall of people.



Mapping Our Anzacs Scrapbook photos viewed through CoolIris

It’s worth noting too that like the markers on the maps, these faces link back to the actual service records. So they’re not just a new way of seeing the collection, they’re a new way of exploring it.

But the records don’t stand in isolation, they themselves have a context. A couple of years ago, Mitchell Whitelaw from the University of Canberra, undertook a project called ‘The Visible Archive‘ to investigate ways of visualising the holdings of the National Archives of Australia. Have you ever wondered what 360km worth of records looks like?



The collections of the NAA visualised by Mitchell's Series Browser.

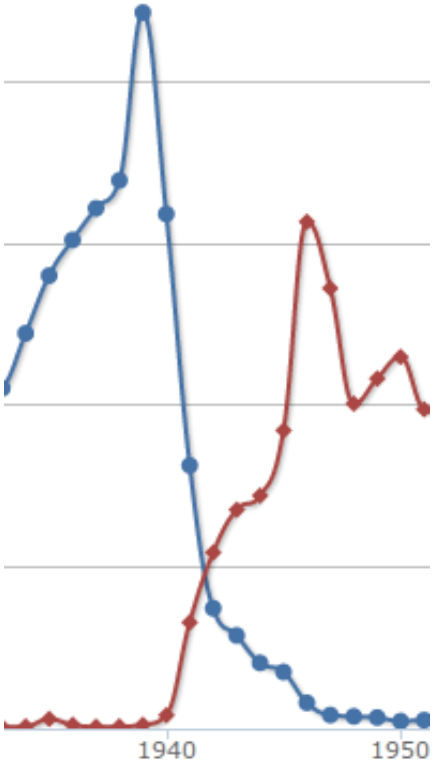
This represents the holdings of the National Archives. Files within the archives are organised into series, and each square in this image represents a single series — there are about 60,000 of them. Naturally the size of the square gives an indication of the size of the series itself. It’s a fascinating and strangely beautiful picture.

It’s easy enough to pick out the World War One service records — Series B2455. In the interactive version of Mitchell’s series browser you can click on a box and display links between series, as well as other series created by the same government agency. Again, it’s not just a way of seeing the collection, but a means of exploring and interpreting it. As Mitchell says:

> Visualisation enables us to literally show everything, to display large volumes of data in a way that reveals patterns and communicates context, but also provides access to the fine grain of individual elements.[1]

But we can also employ such techniques to ask new kinds of questions. Can you imagine how Alexander Kelley and the other inhabitants of the Anzac Hostel must have felt in 1939? They had lost so much in the Great War, the 'war to end all wars', and yet within their own lifetime it was all happening again. More young men were answering the call, more lives were going to be destroyed.

There must have been a dreadful, disheartening moment when Australians realised that the Great War was not an end, but a beginning — the first in a series of devastating global conflicts. At some point the 'Great War' became the 'First World War', but when?



When did the 'Great War' become the 'First World War'?

This is one possible answer. This graph draws its data from the 50 million or so digitised newspaper articles in Trove, the National Library of Australia's discovery service. It shows the proportion of newspaper articles that included the phrase 'the great war' compared to the proportion containing 'the first world war' (and variations thereof). The lines cross late in 1941. With German victories in Europe and Africa, the opening of the Eastern Front and the Japanese attack on Pearl Harbour, 1941 makes sense.

What is perhaps more intriguing is the dramatic peak in the occurrence of 'the great war' in 1939. It's no surprise that the looming threat of a new conflict would provoke comment and comparisons, but it does make you wonder about the context of those discussions and how they might have changed as the reality of war edged closer.

To start exploring this I've harvested the content of the 6,600 articles from 1939 that included the phrase 'the great war'. Using an online text analysis service called VoyeurTools I can quickly generate a picture of their contents.

This simple visualisation shows us the relative frequencies of words within the articles. It doesn't reveal any great mysteries, but it does suggest some possibilities for further prodding. The

prevalence of ‘time’ and ‘new’, for example — might these help us understand the shift in perspective from one war to the next? We can follow this up by browsing the different contexts in which the words were used.

But what actually is it that we’re actually searching? We know that Trove includes newspapers from 1803 to 1954, but if we’re really going to analyse shifting words and ideas it’s important to have a clear picture of the sources of those words.

Something like this perhaps. This graph shows the holdings of the Trove newspaper database on 4 August 2011, organised by state. You can see, for example, that if you’re searching on a topic between the 1920s and 1940s you’re probably likely to get more results from Queensland than anywhere else.

So starting from our location here, today, we can make connections across time and space. We can pull back and look at the big picture, or dive in and examine the fabric of a single life. Through the web we can build and explore a rich and complex contextual network.

It’s an exciting time to be a cultural data hacker. We now have a growing range of tools and technologies available for extracting interesting data from a wide variety of sources, both structured and unstructured.

The ‘Visible Archive’ project started with well-structured data, courtesy of Peter Scott, the developer of the Series System — the descriptive framework used by many Australian archives. But we’re rarely so lucky.

Even when the data starts off in nicely-organised fields in a database there’s no guarantee that that’s how it’s going to be delivered to our web browser. In order to extract the data from my Trove graphs, for example, I had to write a little program called a ‘screen scraper‘ to identify and save the important metadata elements from the raw web page itself.

Where there are no subject keywords we can infer them using techniques such as topic modelling. Where there are no access points we can identify people, organisations, places and events using special tools developed for named entity extraction. Where there are no common identifiers across datasets we can employ record linkage technologies to find possible connections.

We can count words, we can identify parts of speech, we can formulate a measure of the similarity of any two pieces of text. Once we have some useful data we can manipulate and enrich it. Place names can be geolocated — you simply send your place name off to a web service and get back its latitude and longitude.

Increasingly these sorts of tools are becoming accessible to anyone. For historians they offer a means of wrestling with rapidly-growing bulk of source material that is becoming available in digital form. How do you make use of 5 million digitised books, 50 million newspaper articles or the complete archive of every public message ever sent on Twitter?

The digital historian Dan Cohen has noted:

> These computational methods which allow us to find patterns, determine relationships, categorize documents, and extract information from massive corpuses, will form the basis for new tools for research in the humanities and other disciplines in the coming decade.[2]

Dan is involved in a number of interesting projects investigating the possibilities of these

techniques — often grouped together under the heading 'text mining'. One of these projects, 'With Criminal Intent', is looking to see what patterns can be drawn out of the digitised proceedings of criminal trials held at the Old Bailey from 1645 to 1913. That's 197,745 trials, in case you were wondering.

Here's one of their visualisations showing how the length of trials varies over time. Much to the surprise of the research team, this graph suggests a dramatic shift in legal practice around 1825 — defendants started pleading guilty!

A visualisation by the With Criminal Intent project showing changing trial lengths.

Rather than falter under the growing weight of digital sources, these technologies can actually thrive. The more raw material available, the more chance there is to observe and track new patterns. As digitisation continues apace will we ever reach the point when history can simply be read from a graph?

There are some researchers at Harvard who seem to think that's where we're heading. Borrowing liberally from the store of scientific metaphors they have staked out the new field of 'culturomics'. By mining massive digital resources, like Google's scanned books, they hope to map the 'cultural genome' that would enable us to follow the evolution of language and culture.

But there's something quite barren in this ambition. I prefer the vision of digital humanist Stephen Ramsay, who commented in regard to the 'With Criminal Intent' project:

> The Old Bailey, like the Naked City, has eight million stories. Accessing those stories involves understanding trial length, numbers of instances of poisoning, and rates of bigamy. But being stories, they find their more salient expression in the weightier motifs of the human condition: justice, revenge, dishonor, loss, trial. This is what the humanities are about. This is the only reason for an historian to fire up Mathematica or for a student

> trained in French literature to get into Java.[3]

Ultimately it's the stories that nourish, anger, inspire and depress us. The closely-packed map of places recorded in World War I service records is so powerful because we know that under each marker are men, women, families, communities — each with their own story. These new technologies offer new perspectives, they raise new questions, and they challenge us with new contexts to explore and understand. But there is still space for stories and perhaps we can use them to give our stories new life and depth.

This is another World War One service record. It belongs to Charlie Allen. Charlie enlisted three times in the AIF and was discharged on medical grounds each time. It seems he had a problem with his ankle.

Charlie's service record notes a tattoo, proclaiming his love for 'Maud Gordon'. He married Maud in Sydney in 1917 and had two daughters soon after.

Charlie survived the war without further injury, but was not so lucky in peace. On 11 March 1938, Charlie was crushed to death between two railway cars. The accident happened at the Bunnerong Power Station, only a short distance from his home in Matraville. He was buried nearby in the Botany Cemetery.

We also know quite a bit about Charlie's early life. Why? Because Charlie's father was Chinese and he was therefore categorised as a 'half-caste', as someone who was not white, and therefore fell under the restrictions imposed by the White Australia Policy.

Charlie was born in Sydney in 1896. His mother was Frances Allen (sometime sweet shop owner and brothel keeper), his father Charlie Gum (a buyer for Wing On company). Charlie was raised by his mother, but in 1909, at the age of 13, he was taken to China by his father.

Book No. 22

Form No. 91. COMMONWEALTH OF AUSTRALIA. No. 46

DUPLICATE.

CERTIFICATE EXEMPTING FROM DICTATION TEST.

I, John Baxter the Collector of Customs for the State of New South Wales in the said Commonwealth, hereby certify that Charles Albert Jung hereinafter described, who is leaving the Commonwealth temporarily, will be exempted from the provisions of paragraph (a) of Section 3 of the Act if he returns to the Commonwealth within a period of three years from this date.

Date 2 June 09 Collector of Customs.

Nationality Half Caste Chinese Birthplace Sydney NSW

Age 12 years 8 months Complexion

Height Build Medium Hair Dark Eyes Brown

Particular marks Like small scar under right eye

(For impression of hand see back of this document.)

PHOTOGRAPHS.

Full Face. Profile.

Date of departure June 09 Destination China

Ship Eastern

Date of return 5 JUN 1915 Ship Eastern

Port Melbourne

NAA: ST84/1, 1909/22/41-50

This certificate granted Charlie an exemption to the Dictation Test. Without it, he may not have been allowed back into the country.

Every time one of many thousands of non-Europeans resident in Australia sought to travel overseas and return home again they needed one of these certificates.

Charlie's father returned to Sydney, leaving him in China. He lived with relatives in the town of Shekki (inland from Hong Kong). Charlie was naturally homesick, but had no means of getting back to Australia. He wrote to his mother in 1910:

> Do try and bring me home every minute I think of you and long for a piece of bread and butter this tucker is not doing me well.[4]

His mother wrote to the Prime Minister Billy Hughes in an attempt to enlist government help but to no avail. Charlie finally returned to Australia in 1915.

Despite this experience, Charlie visited China again in 1922 for 7 months. Once again carrying papers to grant him re-entry to the country of his birth.

These fragments of Charlie's life have been assembled by my partner, Kate Bagnall, a historian of Chinese-Australia. They are remarkable, and yet not so, because there are many thousands of stories like Charlie's contained within the voluminous records generated by the administration of the White Australia Policy.

We're all of course familiar with the general outlines of the White Australia Policy, and the way it

underpinned conceptions of Australia as a nation in the first half of the 20th century.

But what we sometimes forget is that it was also a massive bureaucratic exercise.

Forms and certificates were printed, issued, used and filed. Regulations were modified, guidelines were distributed and administering officers were managed and advised. Individual cases were reviewed, policy was changed and new forms and certificates were printed, issued, used and filed…

Much of this system is now preserved in the National Archives.

You can get a idea of the range of material available from a case study Kate has prepared focusing on the efforts of Poon Gooey, a successful businessman in Horsham, to keep his wife and family in Australia.

If we look again at Charlie's certificate from 1909 we can see that it contains a lot of interesting structured data:

- name
- place of birth
- age
- height
- destination
- date of departure
- name of ship

We estimate that there are probably about 50,000 of these forms remaining in the Archives, and then there's case files and a variety of other government documents.

Wouldn't it be great if we could extract this structured data. If we could piece together the slivers of identity that remain within the Archives and give people back their lives.

This is the dream of Invisible Australians, a project Kate and I are trying to turn into a reality. Our aim is to build systems that will enable this data to be extracted, aggregated, shared and connected — whether to a family tree, a cemetery record, or another document in another archive.

Imagine being able to navigate the network of lives, families and relationships. To follow their journeys, to share their tragedies, to celebrate their small victories against a repressive system.

Imagine being able to watch them age.

We tend to assume that new technologies require us to change, to adapt. But sometimes they can take advantage of our strengths. Mitchell Whitelaw is interested in finding out what happens when you take large cultural datasets and try to 'show everything'. Such an approach, he suggests, takes advantage of the raw processing power of computers, while giving us space to do what we're good at — finding patterns, making connections, crafting meanings.

The History Wall tries to create a similar sort of space. The History Wall brings together material from a range of different sources — newspaper articles from Trove, biographies from the Australian Dictionary of Biography, records from a database of NSW convicts, population statistics, collection items from the National Museum of Australia — you can pretty much plug anything in as long as it has a date attached to it.

Irish History Wall

For a particular year, the Wall retrieves a random sample from the available sources, jumbles everything up and then throws it onto the screen. As a result, no two views of the Wall are ever quite the same. This is not a traditional exhibition. There is no curator controlling the content or designing the structure. It's ephemeral, it's serendipitous — instead of relying on an authorial voice to smooth over the gaps and transitions, it leaves open the cracks and allows new contexts to seep in and around each item.

As the pioneering digital historian Edward Ayers noted:

> even isolated and inert pieces of evidence — a list, a letter, a map, a picture — can assume new and unimagined meanings when placed in juxtaposition with other fragments.[5]

This is not an absence of narrative, but an opportunity for narration. Edward Ayers suggests that we're actually quite comfortable filling in blanks and untwisting timelines:

> Humans, presented with pieces of information about people, put things into the form of a story. They need not be simple stories, for we know how to deal with unexplained lapses of time, flashbacks, and overlapping narratives. We know how to imagine, infer, things happening at the same time in different places. Film and television train all of us at early ages to weave strands of narrative out of intentional (if carefully constructed) confusion and to take pleasure in that weaving.[6]

And so I can show you a death notice, or a certificate and you will take those fragments, those isolated data points and you will construct a story — you will see the person behind them, you will imagine their life. It's what we do. We're good at it.

Computers on the other hand will just see data.

In her ode in praise of humanities data, digital humanist Amanda French wonders whether we always need to crunch our data into abstract, pliable forms:

> What I wonder is whether instead we can begin with the data, or with a datum, and simply watch for what it may tell us, even if what it tells us is simply a story.[7]

Yes we can. And we should teach computers how to do it as well. Not because we want them to take

over. Not because they can necessarily do it faster or better. But because they can help us share, preserve and connect those stories.

Let's think again about the array of documents that Kate has assembled to piece together the story of Charles Allen. How can you share this sort of material? Typically you'd 'write it up'. You'd capture the story behind the data and commit it to words. The documents would then become evidence — points of connection between your text and the historical record.

So in order to share the meanings of these documents we remove them from the context of the person's life and marshal them as allies to proclaim the authenticity of our rendering. Wouldn't it be better if we could tell the story, but maintain within our texts the direct connections between sources and subject?

What we need is a data framework that sits beneath the text, identifying people, dates and places, and defining relationships between them and our documentary sources. A framework that computers could understand and interpret, so that if they saw something they knew was a placename they could head off and look for other people associated with that place. Instead of just presenting our research we'd be creating a whole series of points of connection, discovery and aggregation.

Sounds a bit far-fetched? Well it's not. We have it already — it's called the Semantic Web.

The Semantic Web exposes the structures that are implicit in our web pages and our texts in ways that computers can understand. The Linked Data movement takes the basic ideas of the Semantic Web and turns them into a collaborative activity. You share vocabularies, so that other people (and computers) know when you're talking about the same sorts of things. You share identifiers, so that other people (and computers) know that you're talking about a specific person, place, object or whatever.

Linked Data is Storytelling 101 for computers. It doesn't have the full richness, complexity and nuance that we invest in our narratives, but it does at least help computers to fit all the bits together in meaningful ways. And if we talk nice to them, then they can apply their newly-acquired interpretative skills to the things that they're already good at — like searching, aggregating, or generating the sorts of big pictures that enable us to explore the contexts of our stories.

This is why we've always imagined Invisible Australians to be something more than an online database. We want to provide points of connection that other people can build into their own stories. But to do that we have to pay attention to things like vocabulary management and authority control, we have to construct web addresses that are not going to break every time we upgrade our software. We have to think about the sorts of things we're talking about — not just people, but government agencies, legislation, certificates, and correspondence. How do we describe these entities and what sorts of relationships do they have?

And of course we need to expose all these structures so that we can say, these things are people, these are events, these are places and these are documents.

Or perhaps, to introduce Alexander Kelley.

Or remember Charles Allen.

You might be wondering why we don't just leave it all to the computers themselves. Didn't I just talk about all the exciting new tools and techniques that enable us to analyse the structures of texts? Perhaps we should just wait for the Culturomics guys to solve all the problems.

But who defines the problems?

Our postmodern sensibilities encourage a suspicion of neutrality. Labels like ‘the new museology’ or Archives 2.0 reflect an awareness that the way we describe and arrange our collections is itself culturally-determined. It’s not just a matter of what our descriptive systems show, but what they hide.

Tim Hitchcock, another member of the ‘With Criminal Intent’ team, has described how online technologies can change the way we access archives. Instead of being forced to navigate the hierarchical structures that archives impose on records, which in turn tend to reflect the workings of the institutions that created the records, we can directly find the people whose lives were regulated, influenced, shaped or controlled by the policies of those institutions.

Instead of merely hearing ‘the institutional voice… in all its stentorian splendour’, he says, we can listen in to ‘the quieter tones uttered by the individual’.[8]

This reminds us that search boxes, along with other digital tools, themselves embody arguments. There are assumptions built into their code about what is relevant, what is significant, what is necessary.

We can build our own tools of course, and we can critique other people’s algorithms. But what if we just want to collect and share stories?

Linked Data gives us a way to present an alternative to Google’s version of the world. We can argue back against the search engines, defining our own criteria for relevance, and building our own discovery networks.

Changing the way we access resources changes the sorts of stories we can tell. Tim Hitchcock asks:

> What happens when institutions and archives are ‘decentred’ in favour of the individual? What changes when we examine the world through the collected fragments of knowledge that we can recover about a single person, reorganised as a biographical narrative, rather than as part of an archival system?[9]

Perhaps the invisible become visible.

1. Mitchell Whitelaw, 'Visualising archival collections: the Visible Archive project', *Archives and Manuscripts*, vol. 37, no. 2, November 2009, pp. 22-40. [↩]
2. Cohen, Dan, 'From Babel to Knowledge', *D-Lib Magazine*, vol. 12, 2006. [↩]
3. Stephen Ramsay, 'Prison Art', delivered at the Digging into Data Conference, June 2011, http://lenz.unl.edu/papers/2011/06/10/prison-art.html [↩]
4. Letter from Charles Allen to Frances Allen, NAA: ST84/1, 1909/22/41-50 [↩]
5. Edward L Ayers, 'History in hypertext', 1999, http://www.vcdh.virginia.edu/Ayers.OAH.html [↩]
6. Edward L Ayers, 'History in hypertext', 1999, http://www.vcdh.virginia.edu/Ayers.OAH.html [↩]
7. Amanda French, 'In Praise of Humanities Data', November 2010, http://www.scribd.com/doc/50066437/In-Praise-of-Humanities-Data [↩]
8. Tim Hitchcock, 'Digital Searching and the Re-formulation of Historical Knowledge', in Mark Greengrass and Lorna Hughes (eds), *The Virtual Representation of the Past*, Ashgate, Farnham, UK, 2008, pp. 81-90. [↩]
9. Tim Hitchcock, 'Digital Searching and the Re-formulation of Historical Knowledge', in Mark Greengrass and Lorna Hughes (eds), *The Virtual Representation of the Past*, Ashgate, Farnham, UK, 2008, p. 90. [↩]

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)

- Share on Facebook (Opens in new window)
- Click to share on Google+ (Opens in new window)
-