Hacking heritage: understanding the limits of online access

Tim Sherratt (University of Canberra)

Preprint of a chapter submitted for publication as part of *The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums and Heritage Sites*.

29 March 2018

Hacking heritage: understanding the limits of online access

Tim Sherratt (University of Canberra)

In 1995, an Australian government plan for digital innovation highlighted some of the exciting possibilities that lay ahead for the cultural sector (Australia, Department of Industry, Science and Tourism, 1995). Access to collections would be 'simplified' through the creation of an 'Electronic Smithsonian' — a portal to bring together the holdings of national cultural institutions:

> *For the user this home page access will be like walking electronically down an avenue of all our major museums or galleries. People will be able to find out about the collections, their significance and context, and use interactive links to other institutions, as well as to access digitalised images.*

Two decades later, the UK government's *Culture White Paper* (UK, Department for Culture, Media and Sport, 2016) envisaged a similar pathway for users, while seeking to make the UK 'one of the world's leading countries for digitised public collections content':

> *We want users to enjoy a seamless experience online, and have the chance to access particular collections in depth as well as search across all collections.*

Digital technologies continue to offer a beguiling vision of universal access. Everyone, everywhere will be able to find and use our cultural collections. Hidden riches will be revealed. Obstacles to discovery and exploration will be removed. Technology, it is often assumed, can push collections across a threshold — it can make them *open*.

But access is never truly open. Voices are suppressed or lost. Information is withheld or restricted. Priorities are set. Technology fails. Yes, putting collections online does create exciting new opportunities for engagement and use, but if we focus on the threshold moments, on the expansion of scope and scale, we draw attention away from the forces that control and shape the cultural record. We make it harder to see what is missing.

This chapter explores what we mean by 'access' in relation to GLAM collections. Instead of seeing the growing wealth of digital collections as a journey towards openness, I want to suggest that digital technologies enable us to critically examine the way access itself is manufactured and controlled.

The promise of online access

There's plenty of evidence that the web, coupled with digitisation, Optical Character Recognition (OCR), and search, has transformed the way we find and use cultural heritage collections. The delivery of historical Australian newspapers through Trove provides one of the most obvious and powerful examples. A user survey in 2013 indicated that use of Trove across Australia generally corresponded with the

national population distribution (Ayres, 2013). People who had never visited the National Library of Australia in Canberra, or even one of the state libraries, are now able to dive deep into their newspaper holdings. Digital delivery has lessened physical isolation. Similarly, the National Archives of Australia has pursued digitisation as a solution to the 'tyranny of distance' (Ling & McLean, 2004). In its 2015-16 Annual Report, the Archives noted that while 111,526 records had been viewed in its reading rooms, 10,579,254 records had been accessed online (National Archives of Australia, 2016).

This is not simply a matter of convenience. People who might never identify as 'researchers', who might never have thought of visiting a major cultural institution, can explore their collections without having to brave the intimidations of architecture or the questioning of gatekeepers, however well-intentioned. Nor is the digital just a replacement for books or microfilms. The application of OCR and full-text search to large text collections such as Trove's digitised newspapers may now seem commonplace, but that doesn't make it any less transformative (Hitchcock, 2008). The change is both quantitative and qualitative. More people are using more resources, but they are also using them differently — navigating patterns, traces, and fragments in a way that would be impossible for even the most hardened microfilm operator (Putnam, 2016).

Using digital technologies GLAM institutions can expose the vast number of collection items that will never make it into physical exhibition. Online collection databases give people the freedom to ask their own questions, to embark on their own adventures of discovery (Cameron & Robinson, 2007). Likewise, institutional authority can give way to new modes of collaboration, as demonstrated by the growing proliferation of online crowdsourcing projects in the cultural heritage domain (Ridge, 2014). More and more institutions are providing openly licensed, and easily downloadable, high-resolution collection images (Kapsalis, 2016). Opportunities are being opened not simply to consume collections as audiences or visitors, but to create with them.

Not too long ago it seemed that every new digital project promised an authoritative 'portal'. Fortunately this time has passed. Curated sets of links quickly fall into neglect and disrepair. Aggregation of digital resources has evolved beyond selection to focus on the possibilities of scale. Services such as Trove, DigitalNZ, the Digital Public Library of America, and Europeana bring together millions of items from GLAM organisations. The purpose of large-scale aggregation is not simply to build better search interfaces, but to develop new platforms for sharing, collaboration, enrichment, and reuse of cultural heritage collections (Sherratt, 2013). Aggregation enlarges the scope and meaning of access.

There is no doubt that digital technologies have changed the way we find and use cultural heritage collections. As the examples above demonstrate, online access has opened GLAM collections to new audiences and new questions. But as the number of online collection items continues to grow, as exciting new interfaces emerge, as more and more organisations share their collection data, it's all too easy to see access as a technology-fuelled march towards some ideal of openness — towards the fabled 'seamless online experience' where the riches of our all cultural institutions are arrayed for easy consumption.

How and why does information become 'open'? And when does it remain 'closed'? By focusing on the technological drivers we obscure the resourcing decisions, the ethical judgements, the political controls, and the historical processes that define the boundary between open and closed and construct our experience of access.

The limits of online access

As Tara Robertson (2018) reminds us, 'not all information wants to be free'. Ethical considerations around privacy and consent should inform decisions about what to digitise. Australian GLAM institutions, for example, generally recognise that access to Indigenous cultural collections should be subject to community consultation and control. The ATSILIRN Protocols, first published in 1995 and updated in 2012, state that access to secret, sacred, or sensitive materials requires careful management in the online environment. Digital knowledge management systems such as Ara Irititja, Keeping Culture, and Mukurtu have all been developed in consultation with Indigenous communities to provide culturally appropriate controls over access. Kimberley Christen (2012), one of the developers of Mukurtu, argues against 'false choices' between open and closed systems, and notes that 'general calls for "open access" undo the social bearings of information circulation and deny human agency'. Access can be withheld for good reasons.

Digitisation itself has a history, rooted not just recent technological development, but in much earlier efforts to expand the reach of access. As Tim Hitchcock (2016) points out, early targets for digitisation were canonical texts microfilmed by commercial firms in the pre-digital age:

> *In other words, what happened in the twentieth century – the aspiration to create a particular kind of universal library, and to commercialise world culture (and to a 1930s mind, this meant male and European culture) – essentially shapes what is now available on line.*

But decisions about what to digitise are only the latest in a series of selections, omissions, erasures, and accidents that have shaped the holdings of our cultural institutions. 'As spaces of power', Rodney Carter (2006) argues, 'the archive is riddled with silences'. Collections are formed by exclusion — by decisions about whose lives, whose voices, matter. Online access is built atop generations of absence and loss. It comes with a responsibility to consider whose experiences are missing from our list of search results. As Lara Putnam (2016) suggests, we need to 'size up the absence'.

The ability to type a few words in a search box and find relevant resources still seems miraculous. But we don't know what we can't find. The apparent omniscience of online discovery systems is maintained by their ability to hide their biases and failures. The reality is different. Safiya Noble (2018) has shown how search algorithms reinforce inequality. Ian Milligan (2013) has pointed at the limitations of OCR when applied to historical texts. Search interfaces lie, OCR is flawed, and metadata is incomplete and inconsistent.

The technological constraints on access become more obvious once you've found an item of interest. How easy is it to copy and share a persistent link to the item? Even after 25 years of the web, software vendors still seem surprised that users would want to do this. Can you download high-resolution copies of images? Can you obtain catalogue data in a machine-readable form for large scale analysis? The meaning of access is closely tied to possibilities for use. Online interfaces construct access by limiting the types of interactions we can have with collections.

Digital technologies give us new freedom to explore collections. It's sometimes claimed that this has resulted in a 'democratisation' of access. More resources are available to more people. But what we see,

and how we see it, are the result of decisions made by someone else. What power do we really have?

In 2017, historian Jenny Hocking took the National Archives of Australia to court. She was arguing for the release of letters from the Governor-General to the Queen relating to the dismissal of the Whitlam government in 1975. However, the court agreed with the Archives, finding that the letters were 'personal' and so not within the definition of Commonwealth Records in the Archives Act. Australian government records are expected to be opened to the public after twenty years. But there are limits — records created by courts and the parliament itself are treated differently, and the Archives Act defines a series of exemptions that can be invoked to withhold records. Access is not inevitable, nor is it guaranteed. Political power, bureaucratic control, and professional practice all play a role in determining what we can see. Hocking's case is interesting because it lifts the lid on some of the processes through which access is negotiated. It turns out that in this case it's ultimately the British Sovereign who gets to decide what we can see.

Online collections have a history. Digital access is the product of analogue processes — of institutional policies, and individual judgments. Our search results are not manufactured by algorithms alone, they are created by many small acts of human imagination, initiative, obstruction, and neglect. But if the development of online access is not an onwards march towards some ideal of 'openness', what is it? How can we track it, analyse it, understand it, or change it?

Many coding languages provide mechanisms for 'introspection'. This allows us to find out something about the properties of a program as it is run. Interfaces to cultural heritage collections offer similar opportunities to look inside, and observe the processes that deliver resources to our browser. Perhaps the simplest example is the number of 'total results' that most search interfaces display. While we can't take this at face value — amongst other things, it's dependent on the configuration of the search index — it does tell us something about the collection as a whole. And that's a place to start.

In the sections that follow, I'll explore how we can use online interfaces against themselves to investigate their own limits. Access in the digital world is not just about what you're given.

Search engines lie

QueryPic (2012) is a tool I created to visualise searches in Trove's digitised newspapers. You enter keywords as you would in the normal web interface, but instead of seeing a list of search results, you see a chart showing the number of matching articles, year by year. QueryPic lets you look for patterns across the complete newspaper corpus.

One QueryPic chart examines the question, when did the 'Great War' become the 'First World War'. It's a good example of how the tool can be used to track changes in language. But if you look closely you'll notice a small bump in the usage of 'First World War' around 1916. How could this be? If you dig down through QueryPic to the relevant articles you'll find that 'First World War' appears not in the newspaper text, but in tags added by Trove users. By default, Trove searches user tags and comments as well as the articles themselves.

Digital access to cultural collections is not just delivered by websites. An increasing number of cultural heritage institutions provide direct access to collection data and images through Application Programming

Interfaces (APIs) or downloadable files. Such data is said to be 'machine-readable' or 'machine-actionable' — instead of being displayed as a product for human consumption, like a web page, it is represented in a form that computers can understand and manipulate. Made available in this form, collection data can support the development of new research, applications, and analyses. The *Collections as Data* (2017) project seeks to develop guidelines, requirements, and use cases for the sharing and reuse of this sort of data.

QueryPic is built using the Trove API. It's a simple example of what becomes possible when cultural institutions make their data available for reuse. But, as the 'First World War' example reveals, machine-readable data inherits the limits and biases of the systems and processes that created it. Trove's search indexes are tuned for easy discovery, not large scale data analysis. There is currently no way to exclude tags and comments from a search using the API. In some ways this compromises the usefulness of QueryPic. It certainly highlights the need for visualisations to be carefully interrogated. But you could also argue that this 'blip' or bug reveals features of the system that are otherwise difficult to see.

Safiya Noble's (2018) analysis of the ways in which search engines reinforce existing prejudices began with a simple Google search for 'black girls'. Matthew Reidsma's (2016) examination of bias in the related topic suggestions of the Summon discovery service began by simply logging user search terms against the topic suggestions. In both cases existing search interfaces were turned upon themselves. Through thoughtful observation and experiment, the interfaces offered up information about their own biases.

The accuracy of OCR is a major issue for services like Trove that offer full-text search across large collections of historical documents. Trove, unlike some commercial services, does at least expose the results of its OCR processing so you can get a feeling for how messy it is. But how does this messiness affect your search results?

Here's a simple experiment to try right now. Go to Trove's digitised newspaper interface and search for 'tbe'. How many results are there? A little browsing will quickly reveal that 'tbe' is a failed OCR attempt at 'the'. When I last tried this search, 'tbe' was found in 14,996,286 articles (about 7% of the total). You could easily repeat this search across particular newspapers, locations, or periods to see how error rates compare.

Search interfaces lie, and they do it with great confidence. Their 'truth' is formulated within the limitations of technology and content. And yet they express no doubt, they offer no qualification. They encourage us to believe that they are comprehensive and accurate. But if we ask the right questions, they can reveal their own limits, sometimes in unexpected ways.

The proceedings of Australia's Commonwealth parliament from 1901 onwards are all available online. The original volumes of Hansard have been scanned, and the text marked up in XML — one file for every sitting day. All this content is searchable through the ParlInfo database, but the interface is not easy to use — for example, it's difficult to simply browse a day's proceedings. In 2016, I wrote a computer script to search ParlInfo and download all of the underlying XML files from 1901 to 1980. The harvested files were shared through a GitHub repository (Sherratt, 2016a), enabling researchers to undertake large scale analysis of political language and events.

But there was a problem. The filenames of some of the XML files I was downloading didn't conform to the standard pattern, and when I looked inside I found they were empty. This meant that some sitting days didn't show up in search results — they were effectively invisible. After more investigation, I discovered that 94 days were missing (Sherratt, 2016b). The gaps were most pronounced in the Senate between the years 1910 and 1919. For example, 21 of 47 sitting days in 1917 weren't appearing in search results. Anyone relying on ParlInfo for research into political responses to the First World War would have missed significant slabs of content. The Parliamentary Library has since replaced the empty XML files, and is undertaking further analysis.

On another occasion, while trying to update details of ASIO surveillance files that I'd previously harvested from the National Archives of Australia's online database, I found that about 400 files had disappeared from the public search interface. The Archives explained that this was an unintended consequence of a recent reorganisation of their holdings.

These things happen. Complex processes fail. The point is not to apportion blame, but to recognise failure as an inevitable part of the experience of online access. While it's unlikely that the missing files would have been noticed by someone using the standard search interfaces of either system, traces remained. These traces tripped up my harvesting programs. Just like the blip in QueryPic, they pointed to anomalies — they helped me to 'size up the absence'.

Search interfaces to cultural heritage collections are not simply services to be consumed, they construct our experience of access. But they also offer glimpses of their inner workings that we can observe and change. What happens when we approach search engines as sites of experimentation and play? Or when we read them as historical documents awaiting close analysis?

Not everything is digitised

What do 200 million newspaper articles look like? Or 300 km of government records? Part of our problem in understanding access is the difficulty of grappling with the scale of our cultural collections.

As Mitchell Whitelaw (2015) argues, the view offered by search boxes is a narrow, miserly slice of our rich collections. He and others have been exploring the possibilities of 'generous interfaces' that encourage exploration by putting collection items up front. Such experiences are no less constructed than a set of search results, and can just as easily deceive their users. But they offer a different approach to the challenges of scale — trusting users to interpret big, abstract pictures, instead of just consuming a stream of bite-sized chunks. Big pictures can prompt us to ask different types of questions, to address the meaning of a collection *as a collection*.

Here's one way of seeing 200 million newspaper articles.

FIGURE 1

This chart simply shows the number of digitised newspaper articles per year in Trove. You can make your own version of this chart using QueryPic — just paste 'http://trove.nla.gov.au/newspaper/result?q=+' into the 'Query url' box, and select the 'number of articles' view. Of course, depending on when you undertake this experiment, your results might be quite different. New articles are being added to Trove all

the time. Has the picture changed?

In 2017, when I constructed this version, there was a significant spike in the number of articles around 1915. Why? Did something notable happen in 1915? No it's not a result of the war… or at least not directly. It's a product of funding and priorities. In the lead up to the centenary of the First World War collaborating libraries decided to focus digitisation dollars on newspapers from the war period. Eventually the spike should flatten out as other gaps are filled, but it's a useful reminder of how online collections are shaped by politics and practicalities.

Creating a big picture of the National Archives' holdings is rather more challenging. Currently, the Archives only provides API access to some First World War service records — once again in support of the Anzac centenary. To obtain any other data from RecordSearch, the Archives' online database, it's necessary to reverse engineer the interface and extract structured information from web pages. This is a process known as screen scraping. Over the years I've scraped large amounts of data and digitised page images from Recordsearch, sharing both the results and the code (Sherratt, 2012, 2016). Screen scraping is inefficient and prone to error, but it's also an example of how access can be negotiated and changed *from the outside* — web pages can be transformed into data; online collections can opened to computational analysis.

In late 2016, I harvested data from 63,711 series in RecordSearch and aggregated information about the numbers of files described and digitised in each series. About a third of all series have at least some item descriptions, and about a fifth have some items digitised. But how is this distributed across the whole collection? Item descriptions make records findable. Digitisation provides instant access. Together they determine the shape and texture of the collection as it is experienced online.

Total series	63,711
Series with item descriptions	20,735 (32.5% of series)
Series with digitised items	4,634 (22.3% of series with descriptions)
Total items described	10,727,214
Total items digitised	1,769,967 (16.5% of items described)

To explore the way access is distributed across the National Archives, I categorised each series using top-level government functions (Sherratt, 2017). These functions are defined in a thesaurus maintained by the Archives, and include things like transport, employment, and immigration. In RecordSearch, functions are associated with government agencies rather than individual series, so there's a fair bit of fuzziness in my groupings. Nonetheless, it's possible to create some collection-level snapshots.

FIGURE 2a and 2b

If we view the subject groupings by the amount of records in each (measured in shelf meters) the distribution seems fairly even. But the picture changes when we focus on the number of items described

or digitised. ‘Defence’ becomes particularly prominent, while areas such as ‘community services’ and ‘cultural affairs’ seem to shrink. The prominence of defence is really no surprise. Service records are heavily used by family historians, and in 2007 the Australian government funded the digitisation of all 375,000 First World War service records in what was branded as ‘A Gift to the Nation’. This one investment has had a significant and enduring impact on access to records held by the National Archives.

There will always be priorities in digitisation programs. There will always be short-term funding opportunities related to specific initiatives or events. There’s nothing wrong with that. It’s just that these biases and distortions are not obvious to someone typing queries into a search box.

Marilyn Lake (2010), Carolyn Holbrook (2017) and others have described how Australian government funding of educational resources related to the First World War has promoted a particular vision of Australian history. Digitisation of war-related collections needs to be examined in this context — what perspectives are privileged by easy online access? The boom in family history has also had an impact on digitisation priorities, encouraging new collaborations with commercial providers like Ancestry. Access to certain collections has a distinct dollar value, a capacity for on-sale to eager genealogists. As Barbara Reed (2014) asks, do organisations understand the compromises they’re making by entering commercial arrangements? Do *we* understand how such priorities affect the range of stories we can tell about the past?

Digitisation shapes our perceptions of reality. The more we have in digital form, the easier cultural heritage collections are to find and use, the more likely we are to assume that everything (or at least everything important) is online. To overcome this, we need to take scale seriously. We need to analyse digitisation priorities in the context of a bigger, unknowable whole.

Access is not always open

The first of January in Australia each year has become an annual celebration of access. Both national and state archives release previously closed files for public scrutiny, and the media fills a slow news day with secrets from governments past. The opening of a new batch of cabinet records by the National Archives of Australia attracts particular attention. But what is generally overlooked in the media coverage is that this is a routine bureaucratic process governed by archives legislation.

With each new year, the ‘open period’ (now set at twenty years) creeps forward, and many more records are potentially available to the public. The cabinet documents are given a head start, pushed through the process of access examination in preparation for their big day. But most records have to wait for a researcher to request them before they are considered for release. Access examination checks the contents of a record against a list of exemptions defined by the Archives Act. These are reasons why files should not be made public — such as privacy or national security. The vast majority of records pass this test and are opened, or partially opened, for all to see. But a small percentage remain closed.

Here’s another experiment to try. Go to RecordSearch on the National Archives website, and select the ‘Advanced search’ for ‘Items’. Go to ‘Access status’ near the bottom of the search form and choose ‘Closed’. RecordSearch will display details of the records you’re not allowed to see.

Since 2016, I’ve been undertaking my own new year’s ritual — harvesting details of all the files in RecordSearch with the access status of ‘closed’. In January 2018 there were 11,235 closed files (Sherratt,

2018b). How many did your search return? I also compile and share lists of files that were newly closed in the past year. It seems only fair that the celebrations surrounding the cabinet release should should spare some time to remember those files that didn't make it through. Access is about what we can't see, as well as what we can.

You might have noticed that RecordSearch not only records the access status of a file, but date the decision was made, and the reasons for the decision. Usually the reasons are specific exemptions defined by the Archives Act, but there are some additional categories. I used the 2016 harvest to create a new interface that lets you look for patterns and connections in a way that is impossible within RecordSearch itself (Sherratt, 2016a). Here, for example, is a chart that shows the number of closed files associated with each 'reason'.

FIGURE 3

The most commonly cited reason is Section 33(1)(g) of the Archives Act which relates to individual privacy. But two other heavily used categories, 'Pre access recorder' and 'Withheld pending adv', are not defined anywhere in the Act. 'Pre access recorder' was used on records that had been closed before the introduction of the Archives Act in 1983. It all but disappeared from the 2017 harvest, as the NAA changed the access status of these files to 'Not yet examined'.

'Withheld pending adv' tells a more complex story. It's used when records are referred back to the government agencies that created or controlled them for advice on whether they can be made public. This process can take months or even years, and so 'Withheld pending adv' is used as a marker to indicate that a file is part way through the examination process. It's closed, but not finally closed. The use of this marker is evidence of a glitch in the system. The Act states that access decisions will be made within 90 days — there is no provision for extended consideration by agencies. This will be changed by proposed amendments to the archives legislation, and so the category might disappear from future harvests.

How closed is 'closed'? Files marked as 'Withheld pending adv' sit in a sort of archival limbo — neither open nor closed. Using data from the 2018 harvest, I was able to identify which of these files were finally released to the public in 2017 and calculate how long they'd been waiting. The average was 3 years and 77 days (Sherratt, 2018b). They might not be finally closed, but they are effectively closed.

In the case of the National Archives of Australia, the meaning of access is defined by legislation. It is assumed that records more than twenty years old will be opened to public scrutiny — the justifications for withholding access are called 'exemptions' for a reason. But that's not the end of the story. Access examination is a complex process involving bureaucratic practice, individual interpretation, and the public right to access. By making regular harvests, I'm hoping to expose this as a historical process — to identify changes over time in the way access is constructed.

The data I'm harvesting is all publicly available through RecordSearch. But the existing interface doesn't allow you aggregate information about sets of files — like Trove, its purpose is discovery not analysis. But when we free the data from the interface new possibilities emerge. *The Real Face of White Australia* made the faces of people living under the White Australia Policy the main point of entry to their files (Sherratt & Bagnall, 2018). Redacted took individual redactions from ASIO surveillance files turned them into a discovery interface — they became a gateway to the files they were intended to censor . One way

of exploring the meaning of access is to turn it inside out. What does access look like when we focus on resources that we're not allowed to see?

Hacking access

It's easy to get excited about the possibilities of online access for cultural heritage collections — new audiences, new uses, new opportunities to demonstrate value and relevance. But there will always be limits. The idea that our efforts are aimed towards a 'seamless' online experience that brings everything together is a dangerous mirage.

In *The Theory and Craft of Digital Preservation*, Trevor Owens (2017) warns 'whatever discovery system or interface you use today is temporary'. Instead of relying on a single point of access, Owens argues for 'multimodal access and use' where collection data is shared in a variety of forms, and new interfaces are created both by GLAM institutions and their users. This is an approach that engages with the complexities and contradictions of access. There is no solution, no off-the-shelf system — all we can do is create, play, build, and critique; to explore access in the making. As Bethany Nowviskie (2016) suggests in her talk on 'speculative collections', there's an opportunity to shift the temporal orientation of our libraries and archives away from a closed and linear past towards an exploration of what might be.

This is not a job for cultural institutions alone. Collection users need to see themselves as more than just the beneficiaries of access. Researchers generally accept that their use of primary source material comes with an obligation to critically engage with its context and meaning. Why should such obligations diminish online? Contexts are multiplied through digitisation, aggregation, and indexing. We should treat interfaces as if they're archaeological sites, digging down through layers of technology, descriptive practice, and institutional history, to understand what is delivered so conveniently through our browsers.

This chapter has provided some examples of how these sorts of excavations might start — from simple experiments using the search box, through to large-scale data harvests, and the creation of new interfaces. This is not intended as a structured research program, but as an invitation to start hacking. Mark Olson (2013) describes a 'hack' as something that 'transforms the effectivities of socio-technical systems, making them work, or *un-work*, often in new and unexpected ways'. Olson explores how adopting a hacker ethos can enlarge the field of humanities practice:

> *a hack can be elegant or kludgy, authored from scratch or patched together and remixed — the important thing is getting things done, pushing the boundaries of what the humanities can do, what effects it can have in the world, and where.*

Hacking the systems that construct and control access to our cultural collections is at the core of humanities practice in the early twenty-first century. As online collections continue to expand, we need to carve out spaces that resist the weight of scale and foster alternative perspectives. As interfaces grow in sophistication and complexity, we need to stage playful and pointed interventions that reveal their limits and empower critique. We don't all have to be coders, but we have to take code seriously. We have to take what we're given and change it.

Hackers might work inside or outside a cultural institution — this is not about us and them. This is about

recognising that for all the resources, intelligence, skill, and care that institutions invest in their online resources they will never be perfect, they will never be finished, they will never be open. Owens, Nowviskie, and others have situated work on digital collections within an ethics of maintenance and care. We are in this for the long haul. Once we give up the dream of universal access we can admit that our systems are broken, and set about the ongoing work of repair.

References

ATSILIRN. (2012). *Aboriginal and Torres Strait Islander Protocols for Libraries, Archives and Information Services*. Retrieved from http://atsilirn.aiatsis.gov.au/protocols.php

Ayres, M. (2013, July). *Singing for their supper: Trove, Australian newspapers, and the crowd*. Presented at the IFLA WLIC 2013, Singapore. Retrieved from http://library.ifla.org/245/1/153-ayres-en.pdf

Cameron, F., & Robinson, H. (2007). Digital Knowledgescapes: Cultural, Theoretical, Practical, and Usage Issues Facing Museum Collection Databases in a Digital Epoch. In F. Cameron & S. Kenderdine (Eds.), *Theorizing Digital Cultural Heritage: A Critical Discourse* (pp. 165–191). MIT Press.

Carter, R. G. S. (2006). Of Things Said and Unsaid: Power, Archival Silences, and Power in Silence. *Archivaria*, 61(Spring), 215–233. Retrieved from https://archivaria.ca/index.php/archivaria/article/view/12541

Christen, K. (2012). Does information really want to be free? Indigenous knowledge systems and the question of openness. *International Journal of Communication*, 6. Retrieved from http://ijoc.org/index.php/ijoc/article/view/1618

Collections as Data. (2017). Retrieved from https://collectionsasdata.github.io/

Department of Industry, Science and Tourism. (1995). *Innovate Australia*. Retrieved from http://webarchive.nla.gov.au/gov/19961102131909/http://www.dist.gov.au/events/innovate/itt.html

Department for Culture, Media and Sport. (2016). *The Culture White Paper*. Retrieved from https://www.gov.uk/government/uploads/system/uploads/attachment*data/file/510798/DCMS*The*Culture* White*Paper**3*.pdf

Hitchcock, T. (2008). Digital searching and the re-formulation of historical knowledge. In Mark Greengrass & Lorna Hughes (Eds.), *The Virtual Representation of the Past* (pp. 81–90). Farnham, UK: Ashgate.

Hitchcock, T. (2016, July 6). The Digital Humanities in Three Dimensions. Retrieved from http://historyonics.blogspot.com.au/2016/07/the-digital-humanities-in-three.html

Hocking, J. (2017, September 6). The palace treats Australia as the colonial child not to be trusted with its own history. *The Guardian*. Retrieved from http://www.theguardian.com/commentisfree/2017/sep/06/the-palace-treats-us-as-the-colonial-child-not-to-be-trusted-with-knowledge-of-our-own-history

Holbrook, C. (n.d.). Adaptable Anzac: Past, present and future. In Honest History (pp. 48–63). Sydney: New South.

Kapsalis, E. (2016). *The impact of Open Access on galleries, libraries, museums, & archives*. Smithsonian Emerging Leaders Development Program. Retrieved from http://siarchives.si.edu/sites/default/files/pdfs/2016*03*10*OpenCollections*Public.pdf

Lake, M. (2010). How do schoolchildren learn about the spirit of Anzac? In M. Lake & H. Reynolds (Eds.), What's wrong with Anzac: the militarisation of Australian history (pp. 135–156). Sydney: New South.

Ling, T., & McLean, A. (2004). Taking it to the People: Why the National Archives of Australia Embraced Digitisation on Demand. *Australian Academic & Research Libraries*, 35(1), 2–15. https://doi.org/10.1080/00048623.2004.10755253

Milligan, I. (2013). Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History, 1997–2010. *Canadian Historical Review*, 94(4), 540–569. https://doi.org/10.3138/chr.694

National Archives of Australia. (n.d.). *Access to records under the Archives Act - Fact sheet 10*. Retrieved from http://naa.gov.au/collection/fact-sheets/fs10.aspx

National Archives of Australia. (2016). *Annual Report 2015-16*. Retrieved from http://www.naa.gov.au/about-us/publications/annual-reports/2015-16/index.aspx

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Nowviskie, B. (2016, October 27). Speculative collections. Retrieved from http://nowviskie.org/2016/speculative-collections/

Olson, M. J. (2013). Hacking the humanities: Twenty-first-Century literacies and the 'becoming other' of the humanities. In E. Belfiore & A. Upchurch (Eds.), *Humanities in the Twenty-first Century: Beyond utility and markets* (pp. 237–250). Palgrave Macmillan.

Owens, T. (2017). *The Theory and Craft of Digital Preservation*. LIS Scholarship Archive. https://doi.org/10.17605/OSF.IO/5CPJT

Putnam, L. (2016). The Transnational and the Text-Searchable: Digitized Sources and the Shadows They Cast. *The American Historical Review*, 121(2), 377–402. https://doi.org/10.1093/ahr/121.2.377

Reed, B. (2014). Reinventing access. *Archives and Manuscripts*, 42(2), 123–132. https://doi.org/10.1080/01576895.2014.926823

Reidsma, M. (2016, March 11). Algorithmic Bias in Library Discovery Systems. Retrieved from https://matthew.reidsrow.com/articles/173

Ridge, M. (Ed.). (2014). *Crowdsourcing our Cultural Heritage*. Ashgate Publishing, Ltd.

Robertson, T. (2018). Not All Information Wants to be Free: The Case Study of On Our Backs. In P. D. Fernandez & K. Tilton (Eds.), *Applying Library Values to Emerging Technology: Decision-Making in the Age of Open Access, Maker Spaces, and the Ever-Changing Library* (Publications in Librarianship #72) (pp. 225–239). American Library Association. Retrieved from http://eprints.rclis.org/32463/

Sherratt, T. (2012a). *QueryPic*. Retrieved from http://dhistory.org/querypic/

Sherratt, T. (2012b). recordsearch_tools. Retrieved from https://github.com/wragge/recordsearch_tools

Sherratt, T. (2013, October). *From portals to platforms: building new frameworks for user engagement*. Presented at the LIANZA 2013 Conference, Hamilton, New Zealand. Retrieved from http://www.nla.gov.au/our-publications/staff-papers/from-portal-to-platform

Sherratt, T. (2015). Asking better questions: History, Trove and the risks that count. In Phillipa McGuinness (Ed.), *Copyfight* (pp. 112–124). NewSouth Publishing. Retrieved from http://discontents.com.au/asking-better-questions-history-trove-and-the-risks-that-count/

Sherratt, T. (2016a). *Closed access*. Paper presented at DH2015, Hobart. Retrieved from http://discontents.com.au/closed-access/

Sherratt, T. (2016b). hansard-xml. Retrieved from https://github.com/wragge/hansard-xml

Sherratt, T. (2016c). Investigating the Hansard black hole. Retrieved from http://timsherratt.org/research-notebook/historic-hansard/notes/investigating-the-hansard-black-hole/

Sherratt, T. (2016d). recordsearch-series-harvests. Retrieved from https://github.com/wragge/recordsearch-series-harvests

Sherratt, T. (2016e). redacted. Retrieved from http://owebrowse.herokuapp.com/redactions/

Sherratt, T. (2017). Viewing the NAA through functions. Retrieved from http://timsherratt.org/research-notebook/aggregated-archives/notes/naa-functions-view/

Sherratt, T., & Bagnall, K. (2018, in press). The People Inside. In K. Kee & Compeau, Tim (Eds.), *Seeing the Past: Experiments with Computer Vision and Augmented Reality in History*. University of Michigan Press.

Sherratt, T. (2018, February 2). Withheld, pending advice. *Inside Story*. Retrieved from http://insidestory.org.au/withheld-pending-advice/

Whitelaw, M. (2015). Generous Interfaces for Digital Cultural Collections. *Digital Humanities Quarterly*, 9(1). Retrieved from http://www.digitalhumanities.org/dhq/vol/9/1/000205/000205.html

FIGURE1

Caption: Number of digitised newspaper articles available in Trove by year

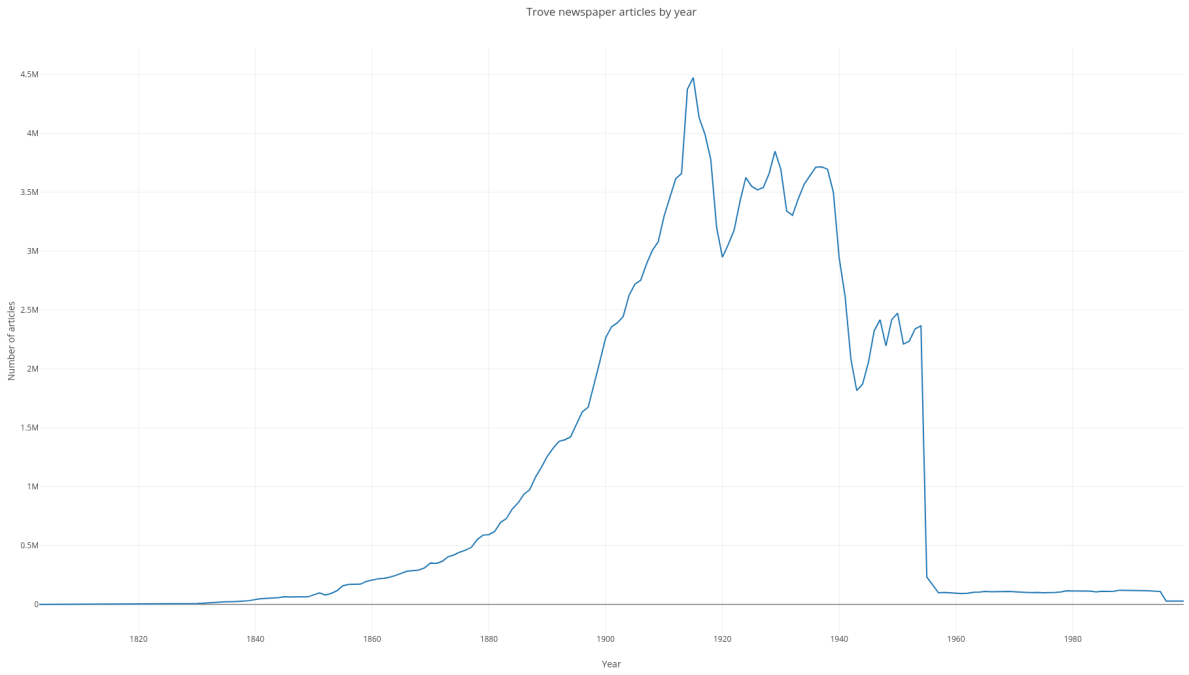

FIGURE 2a

Caption: Top-level functions in the National Archives of Australia by quantity (shelf metres)

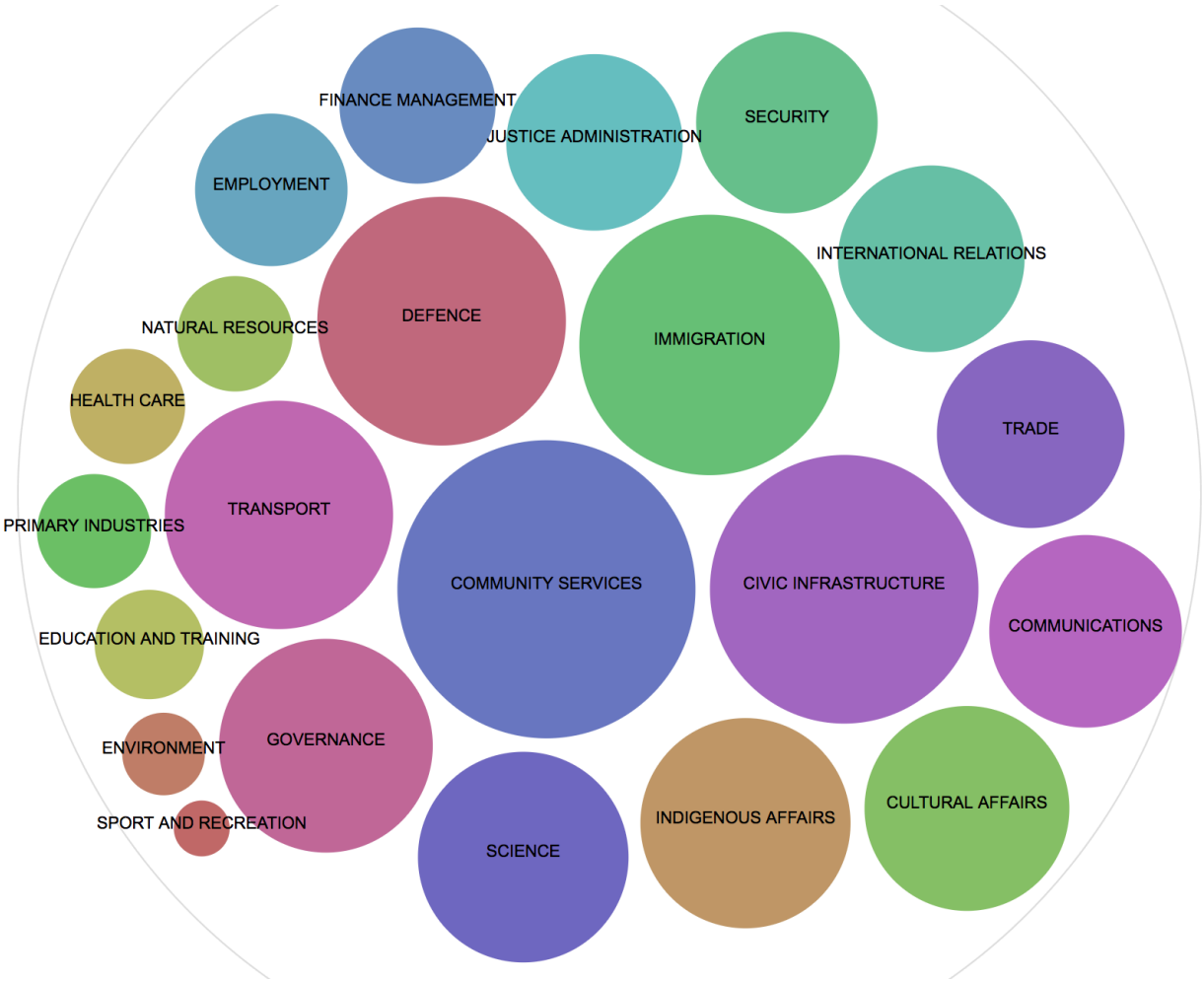

FIGURE 2b

Caption: Top-level functions in the National Archives of Australia by number of items digitised

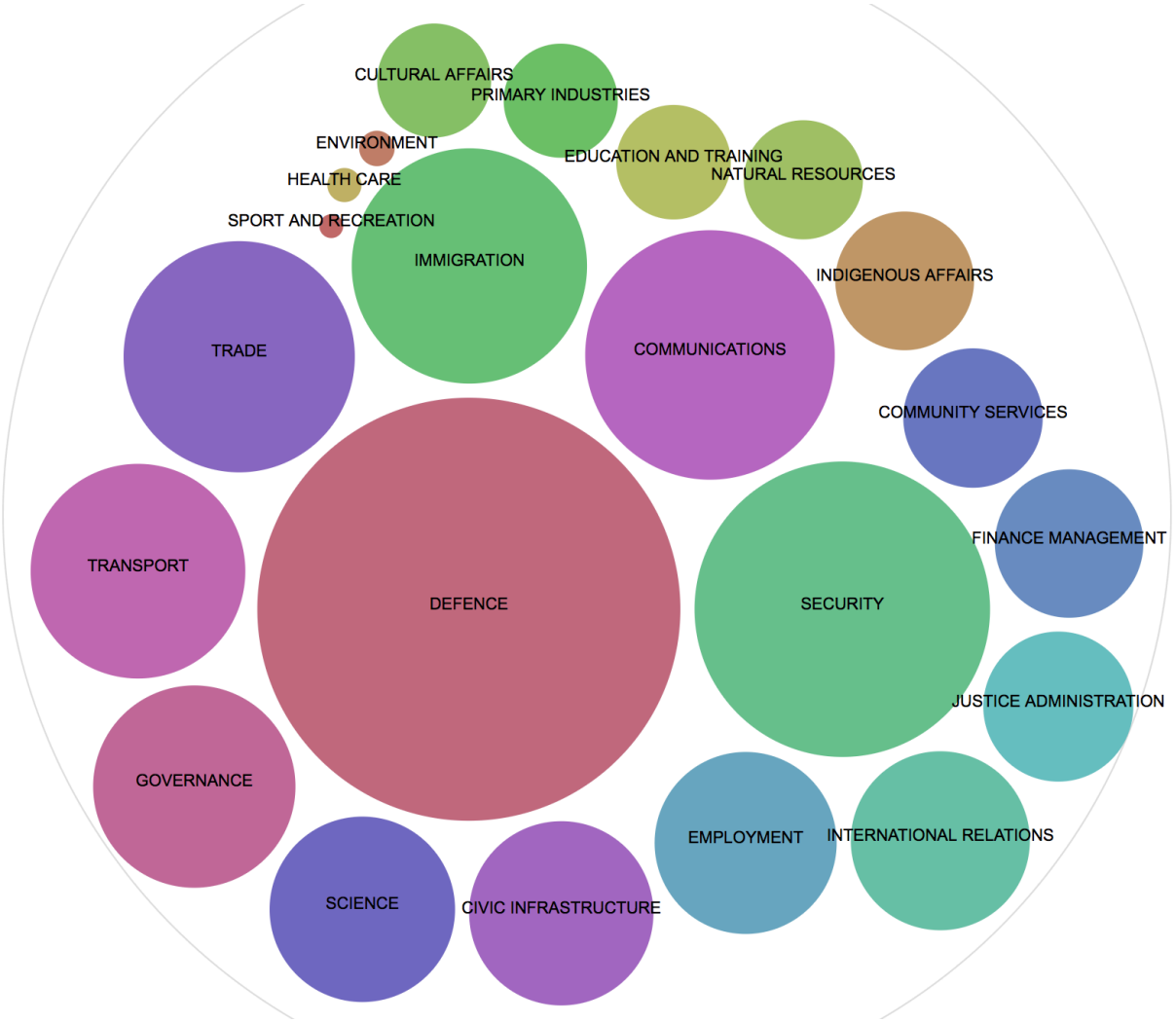

FIGURE 3

Caption: Reasons cited for closing files in the National Archives of Australia