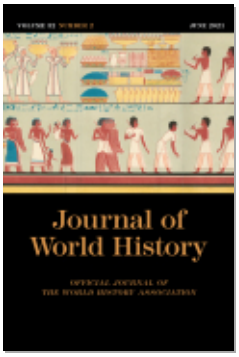
Missing Links: Data Stories from the Archive of British Settler Colonial Citizenship



Kate Bagnall, Tim Sherratt

Journal of World History, Volume 32, Number 2, June 2021, pp. 281-300 (Article)

Published by University of Hawai'i Press
DOI: https://doi.org/10.1353/jwh.2021.0025

Missing Links: Data Stories from the Archive of British Settler Colonial Citizenship

KATE BAGNALL AND TIM SHERRATT

Digitized sources and digital methods are changing the way that we do history. For historians of the British Empire, the digital age offers new possibilities for investigating the lives of those who moved around the empire and across the world. However, much discussion of the possibilities and problems of digital history have focused on the creation and use of full text resources, skipping over the analytical opportunities offered by the descriptive systems in which those texts are embedded. This article is an attempt to fill this gap by documenting a journey through archival data relating to nineteenth-century Chinese naturalization in the Pacific Rim settler colonies of Australia, New Zealand, and Canada. We argue that such data stories are critical if we are to understand both possibilities and pitfalls of research in digital collections.

KEYWORDS: archives, access, historical data, digitization, naturalization, citizenship, British Empire.

DIGITIZED sources and digital methods are changing the way that we do history. For historians of the British settler colonies, the digital age offers particular possibilities for investigating the lives of those who moved around the empire and across the world, from English convicts sentenced to transportation to Cantonese migrants searching for new opportunities. Such investigations can be undertaken with a different scope and at a faster pace than in the analogue past, as online finding aids such as collection databases and name indexes, as well as digitized archives themselves, make it possible to both piece together detailed microhistories and undertake large-scale studies using "big data."

Journal of World History, Vol. 32, No. 2

Much discussion of the possibilities and problems of history in the digital age has rightly focused on the creation and use of full-text resources. The extraction of text from digital images through transcription, Optical Character Recognition (OCR), crowdsourcing, and now handwritten text recognition opens the content of archival collections to discovery, manipulation, analysis, and enrichment. Large-scale projects like the Old Bailey Online and the Digital Panopticon have shown how the extraction of machine processable data from textual sources supports a wide range of research questions and methodologies.[1] And a growing number of "small histories" of individual lives have also demonstrated the expedience of online resources for discovery and connection.[2] But if, as Verne Harris argues, archival collections represent just "a sliver of a sliver of a sliver" of the documentary record, we must recognize that the selection and prioritization of investments in digitization and full-text extraction create ever narrower slices of the past.[3]

In focusing on digital texts, historians have tended to skip over the analytical opportunities offered by the descriptive systems in which those texts are embedded. Finding aids, item lists, catalogues, collection databases, and indexes seem cumbersome in comparison to full-text search, but they cast a wider net over collections. Before Named Entity Recognition (NER) or Natural Language Processing (NLP) were used to identify features in texts, archivists and librarians were creating name indexes and subject lists, documenting relationships among records and creators, and building a range of useful access points. These descriptive systems can themselves be opened to new computational processes, allowing historians to look for patterns and connections that cannot be observed through existing interfaces. For example, Tim's "Closed Access" project analyses the details of "closed"

[1] See Tim Hitchcock, Robert Shoemaker, Clive Emsley, Sharon Howard, and Jamie McLaughlin et al., *The Old Bailey Proceedings Online, 1674–1913*, accessed March 14, 2020, http://www.oldbaileyonline.org, and *The Digital Panopticon: Tracing London Convicts in Britain and Australia, 1780–1925*, accessed March 14, 2020, http://www.digitalpanopticon.org.

[2] See, for example, Julia Laite, "The Emmet's Inch: Small History in a Digital Age," *Journal of Social History* shy118 (published online February 2019; print forthcoming), doi:10.1093/jsh/shy118; Clare Anderson, *Subaltern Lives: Biographies of Colonialism in the Indian Ocean World, 1790–1920* (Cambridge, England: Cambridge University Press, 2012); Kate Bagnall, "'To His Home at Jembaicumbene': Women's Cross-Cultural Encounters on a Colonial Goldfield," in *Migrant Cross-Cultural Encounters in Asia and the* Pacific, ed. Jacqueline Leckie, Angela McCarthy, and Angela Wanhalla (Abingdon & New York: Routledge, 2017), 56–75.

[3] Verne Harris, "The Archival Sliver: Power, Memory, and Archives in South Africa," *Archival Science* 2 (March 2002): 63–86, doi:10.1007/BF02435631.

files in the National Archives of Australia to reveal the workings of the access examination process.[4]

However, getting the data can be difficult. Most of the systems we work with do not provide direct access to the collection data underpinning their online interfaces. Indeed, a significant proportion of this article will be devoted to documenting the problems, caveats, and workarounds associated with obtaining useful data. Initiatives such as the "Collections as Data" project hope to improve this situation by raising awareness of the value of collection data, and supporting cultural heritage institutions in the development and sharing of datasets that "support computationally driven research and teaching."[5] But with a proliferation of systems and standards layered on top of the idiosyncrasies of the historical record, we cannot expect this to be easy. There will always be work to do.

We argue that this is work for historians as well as libraries, archives, and museums. Despite critical engagement by historians with the content and organization of archives,[6] there has been little examination of technologies through which digital collections are found and used. Jennifer Guiliano argues that as historians "we must recognize the digital, and its multiplicity of forms, as historical objects that are produced, interpreted, and contested."[7] She exhorts historians to "encounter the computer and the digital with a skeptical eye" and "actively work to understand how decisions about design, programming, expression, interface, networking, access, sustainability, etc. produce and privilege certain types of history." The systems we use to access archives online embed assumptions about scope, meaning, and significance that we should be prepared to unpack. Tim has suggested that humanities researchers should treat collection interfaces as archaeological sites, "digging down through layers of technology, descriptive practice, and institutional history, to understand what is delivered so conveniently through our browsers."[8] Jim Mussell, in discussing "history as digital practice," suggests treating the archive

[4] Tim Sherratt, "Withheld, Pending Advice," *Inside Story* (2 February 2018), https://insidestory.org.au/withheld-pending-advice/.

[5] See Thomas Padilla and Laurie Allen et al., *Always Already Computational: Collections as Data*, accessed March 14, 2020, https://collectionsasdata.github.io.

[6] See, for example, Kirsty Reid and Fiona Paisley, eds., *Sources and Methods in Histories of Colonialism: Approaching the Imperial Archive* (London and New York: Routledge, 2017).

[7] Jennifer Guiliano, "Towards a Praxis of Critical Digital Sport History," *Journal of Sport History* 44, no. 2 (2017), 146–159, preprint, http://hdl.handle.net/1805/15839.

[8] Tim Sherratt, "Hacking Heritage: Understanding the Limits of Online Access," in *The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums and Heritage Sites*, ed. H. Lewi et al. (Routledge, 2019).

itself as a "large dataset" for critical analysis and visualization, and, redeploying a familiar archives trope, proposes historians use "digital resources against the grain of their interfaces in order to access the data they contain."[9]

Digital research is beset by missing links, both literal and figurative. What data is available? Where is the documentation? What skills do you need? What tools and technologies are available? This article is an attempt to fill in some of these gaps by documenting a journey through archival data relating to naturalization in the British settler colonies of Australia, New Zealand, and Canada over the nineteenth and early twentieth centuries—what we have termed "the archive of British settler colonial citizenship." Our focus is deliberately on processes rather than outcomes. Historians rarely talk about the realities of archival research—the work involved in interpreting finding aids, deciphering file references, understanding recordkeeping systems, or navigating the rules of an unfamiliar reading room—yet this complex set of skills and practices shapes the nature of historical research. Similarly, to work with archival collections as data, a researcher first needs to understand what sort of data is available and how can it be accessed. This is rarely straightforward and frequently frustrating. Guiliano, Mussell and others writing about the practice of digital history have called for exploratory methodologies to be documented and shared.[10] We argue that the analysis of collection data will enrich the scope and context of research in digital history, but to achieve this we need to document the methodological challenges in finding, using, and interpreting such data.

The data explorations we document in this article are part of a larger comparative and transnational project examining Chinese naturalization in colonial New South Wales (NSW), New Zealand (NZ), and British Columbia (BC). Each of the archives that holds records of particular relevance to the broader project—NSW State Archives, National Archives of Australia (NAA), Archives NZ, BC Archives, and Library and Archives Canada (LAC)—has made substantial collection information available online, including name-based indexes and, in some instances, digitized records. Yet the available information varies and the access is patchy, meaning that it has been necessary for

[9] Jim Mussell, "Doing and Making: History as Digital Practice," in *History in the Digital Age*, ed. Toni Weller (Routledge, 2013), 79–94.

[10] Fred Gibbs and Trevor Owens, "Hermeneutics of Data and Historical Writing," in *Writing History in the Digital Age*, ed. Kristen Nawrotzki and Jack Dougherty (Ann Arbor, MI: University of Michigan Press, 2013), https://hdl.handle.net/2027/fulcrum.xw42n885n.

Kate to travel for extended periods to access physical archives in Sydney, Canberra, Wellington, Victoria BC, and Ottawa. Working on a transnational project that sits firmly in the global context of the British Empire, and with records spread across at least five archival institutions in three countries, we were prompted to think about how the information archives make available online might be garnered and analyzed using digital methods. The broader project, like many others, involves the laborious compilation and analysis of specific historical data, but what about the general archival data that is already there?[11] Could it perhaps be a way into a big picture analysis of numbers, processes and patterns of naturalization across the British settler colonies?

Our aim has been to explore how such an analysis might start, by bringing together datasets from archival organizations in Australia, New Zealand, and Canada that relate to individual applications for naturalization. Ultimately, we hope to use the data to observe change over time, the impact of legislation, and differences in ethnic or national origin. But first we need to know what data is available. We have focused on three types of naturalization records and three sources of archival data. The records are: applications for naturalization and their corresponding paperwork; certificates of naturalization; and registers and indexes of naturalizations. The sources of archival data are:

- online archival finding aids, primarily collection databases such as the National Archives of Australia's RecordSearch and Archives NZ's Archway;
- online name indexes, created and maintained by archival organizations and/or third parties such as genealogical societies and Ancestry.com;
- digitized records, created and maintained by archival organizations and/or third parties such as FamilySearch and Ancestry.com.

This article documents our efforts first to access this data, and then use it to explore how the number of naturalizations changed over time in each jurisdiction. Our approach is critical and experimental—we are not simply finding data, we are trying to document its characteristics and history. As becomes clear, even these modest aims require us to grapple with the nature of access, the transformation of data, and the contexts of data creation and use.

[11] Part of the broader project is the compilation of biographical databases of Chinese who were naturalized in New South Wales (to 1888, when Chinese naturalization was prohibited by law), New Zealand (to 1908, when it was prohibited by Cabinet decision), and British Columbia (to 1914, when new legislation and administrative processes were introduced).

Naturalization and its Archival Legacy

The nineteenth century was, in legal historian Helen Irving's words, "the hinge between a world where citizenship meant relatively little and a world in which it was profoundly important to the fate of individuals."[12] For the British colonies of the Pacific Rim, the nineteenth century was also a time of unprecedented numbers of immigrant "settlers" arriving not only from the United Kingdom, but also from Europe, Asia, and the United States. In Australia, New Zealand, and Canada, as across the British Empire as a whole, the legal categories of "subject" and "alien" defined the status of both immigrant and native-born people. Following the common law principle of *jus soli*, anyone born on British soil was a British subject, whether that was in London or in a distant corner of the empire, while everyone else was an alien.[13] Naturalization was the legal process through which alien immigrants could become British subjects when they settled in British territory.[14]

Until the introduction of the *Aliens Act 1844* (UK), naturalization of foreigners had been accomplished by individual Acts of Parliament.[15] However, the new 1844 Act introduced a more straightforward, administrative form of naturalization, which then became the model for corresponding colonial laws. Beginning with South Australia in 1846, by 1871 each of the Australasian colonies, British Columbia, and the Dominion of Canada had enacted its own naturalization legislation, primarily aimed at easing the settlement of "alien friends."[16] Over the passing decades of the late nineteenth

[12] Helen Irving, *Citizenship, Alienage, and the Modern Constitutional State: A Gendered History* (Cambridge: University of Cambridge Press, 2016), 49.

[13] The only exceptions were children born to foreigners during a hostile occupation and the children of foreign ambassadors. Alexander Cockburn, *Nationality; or, The Law Relating to Subjects and Aliens Considered with a View to Further Legislation* (London: William Ridgway, 1869), 7–12, https://archive.org/details/cu31924052577834.

[14] On the history of British nationality across the empire, see Rieko Karatani, *Defining British Citizenship: Empire, Commonwealth and Modern Britain* (London and New York: Routledge, 2014).

[15] A separate but related process, denization, could grant an alien some of the rights of a British subject, particularly the right to own land. New South Wales and Tasmania (then known as Van Diemen's Land) enacted denization legislation in 1828 (9 Geo IV n 6) and 1834 (5 Wil IV n 4), respectively.

[16] The first local naturalization law was introduced in: South Australia in 1846 (20 of 21 Vic); New South Wales in 1847 (granted royal assent in 1849) (11 Vic n 39); British Columbia in 1859 (*Aliens Act 1859*); Vancouver Island (*Alien Act 1861*); Tasmania (25 Vic n 2) and Queensland in 1861 (25 Vic n 9); Victoria in 1863 (26 Vic n 166); New Zealand in 1866 (30 Vic n 17); Dominion of Canada in 1868 (31 Vic c 66); and Western Australia in 1871 (35 Vic n 2).

century, these colonial naturalization laws were refined and localized, in response to changes in imperial law and to particular local social and economic conditions.[17]

The Pacific Rim settler colonies were immigrant societies, and naturalization was a primary way that colonial governments shaped the make-up of their growing populations. As noted, the introduction of naturalization legislation across the colonies was largely done with the aim of turning European immigrants, be they German or Swiss or Swedish or American, into settlers; naturalization was a way of including these alien immigrants in the colonial project as land owners, workers, and voters. Legislation, regulations, and administrative processes were therefore adapted over time as a tool of migrant inclusion and exclusion. In New Zealand, for example, the fee for naturalization was initially set at one pound in 1866, but to encourage naturalization this was reduced in 1882 to two shillings and sixpence, and then abolished altogether in 1892—except in the case of Chinese, for whom the initial high fee remained in place throughout.[18] Similarly in Queensland, the *Aliens Act 1867* set out different requirements for the naturalization of "Asiatic and African aliens" and "European and North American aliens," including that Asiatic and African applicants must be married, have lived in the colony for three years, and have their wife living with them at the time of application.[19] The Chinese, who were the largest nonwhite immigrant population in colonial Australia, New Zealand, and British Columbia, were the group who primarily encountered the exclusionary power of naturalization, both by specific laws and regulations as noted above, or through administrative decisions.

The history of naturalization law in the British Empire provides the framework for our study of the archive of settler colonial citizenship. Following on from legislation were regulations, administrative processes, bureaucratic decisions, legal precedents, and documentation in the form of applications, oaths, correspondence, certificates, and registers. The law itself often set out the fundamentals of these

[17] Between 1844 and 1923, the nine colonies and three dominions under discussion (see footnote 16) enacted more than 40 principal and amendment Acts concerning aliens and naturalization. For example, Victoria, which became a separate Crown colony in 1851, amended its law relating to aliens three times following its introduction in 1863—in 1865, 1890, and finally in 1896; from Federation in 1901 the "aliens power" transferred to the Commonwealth of Australia.

[18] *Aliens Act 1866* (NZ) (30 Vic n 17); *Aliens Act Amendment Act 1882* (NZ) (46 Vic n 17); *Aliens Act Amendment Act 1892* (NZ) (56 Vic n 19).

[19] *Aliens Act 1867* (Qld) (31 Vic n 28).

processes, including paperwork to be submitted and records to be kept. The New South Wales *Aliens Act 1847*, for example, stated that the applicant (known as the memorialist) must present a written memorial requesting naturalization to the Governor, who would then decide to grant a certificate or not. If granted, the memorialist would swear an oath of allegiance and pay a fee, and the certificate of naturalization would be registered with, and a copy kept by, the Supreme Court of New South Wales. Within a common imperial framework, there were obvious similarities in the legal and bureaucratic systems used to administer naturalization across the colonies, but importantly they were not exactly the same. For example, in British Columbia (and Canada more broadly) the decision to grant naturalization rested with the provincial courts rather than with the Governor as in the Australasian colonies.[20] Other differences in law, administration, and recordkeeping came with jurisdictional changes as colonies were established, merged, and separated, and then as colonies became dominions. The differences between jurisdictions and over time are reflected in the archival legacies we have to work with as historians today.

Access Versus Use

Archival data comes in many forms, and the first step in making use of it is to understand the types of access that different institutions and interfaces afford. Online collection databases and finding aids typically provide descriptive information about an archival collection organized in a structured form—often called metadata. The structure is important as it enables analysis and aggregation of the metadata according to particular values, such as the date or the creator. Think of the difference between a spreadsheet and a narrative—both can describe a collection but the spreadsheet allows you to focus on specific pieces of information. The National Archives of Australia and Archives NZ maintain large collection databases, RecordSearch and Archway, that provide structured information about millions of resources. Based on the series system, these databases document complex relationships between entities such as agencies, functions, series, and items.[21] For example, a government agency might be identified as creator of a

[20] Compare, for example, *Aliens and Naturalization Act 1868* (Canada) (31 Vic c 66), with *Aliens Act 1866* (NZ) (30 Vic n 17) and *Aliens Act 1861* (Tas) (25 Vic n 2).

[21] For more on the series system, see Adrian Cunningham, Laura Millar, and Barbara Reed, "Peter J. Scott and the Australian 'Series' System: Its Origins, Features, Rationale,

record series, which itself is identified as the parent of an individual file. RecordSearch and Archway, and other archival collection databases, are available online and can be browsed and searched by users. But do they provide useful data?

Online collection databases do not necessarily deliver their data in a form that can easily be analyzed. Humans can usually interpret a list of search results on a web page without much help. But if computer programs are to make use of search results, they need explicit markers to identify individual fields and records; to understand the difference between a field's label and its value; and to know whether a value should be represented as a number, a date, or text. Data that is clearly structured and delivered in a standard format is said to be "machine readable." BC Archives, for example, which is part of the Royal British Columbia Museum, uses a collection management system that makes descriptions available in standard, machine readable formats such as EAD and EAC-CPF.[22] Despite this, we encountered system errors that created difficulties in downloading details of BC county court naturalization applications. Eventually, through a semimanual process, we were able to obtain CSV formatted file lists. CSV (comma separated values) files are often used to share tabular data in a machine readable form. They can be opened like a spreadsheet and manipulated by a variety of programs for qualitative or quantitative analysis.

Neither RecordSearch nor Archway provide machine readable data. Their interfaces are designed to support discovery by humans, rather than analysis by machines. To extract machine readable data from these systems, we have to understand how they work. Specifically, we need to know how to navigate our way through a complete set of search results, and how to identify the data we want within the HTML code of each page. Once we have this information, we can write a computer program that turns web pages into structured data. This is a process known as "screen scraping." Tim has developed and shared screen scrapers for use with RecordSearch and Archway, which make it possible to generate machine readable datasets from online searches. [23]

Impact and Continuing Relevance," *Comma* 2013, no. 1 (2013): 121–144, https://doi.org/10.3828/comma.2013.1.13.

[22] EAD (Encoded Archival Description) and EAC-CPF (Encoded Archival Context – Corporate Bodies, Persons and Families) are archival descriptive standards. See Encoded Archival Description, Technical Subcommittee for Encoded Archival Standards of the Society of American Archivists, accessed March 14, 2020, https://www.loc.gov/ead/.

[23] See Tim Sherratt, GLAM-Workbench/recordsearch (Version v0.1.0), Zenodo, accessed March 14, 2020, http://doi.org/10.5281/zenodo.3544754, and GLAM-Workbench/archway-harvesting (Version v0.1.0), Zenodo, accessed March 14, 2020, http://doi.org/10.5281/zenodo.3544700.

For the purposes of this study we have, for example, harvested details from NAA series A711, which contains South Australian memorials of naturalization dating from 1865 and 1903, and have shared the list of files through our project's repository as a CSV formatted text file.

Using screen scraping to compile data can, however, be frustrating and unreliable. RecordSearch and Archway both impose limits on the number of results returned by a search, presumably for performance reasons (that is, the speed at which results are returned). This is unlikely to cause a problem to a human user trying to focus in on a particular set of files, but it hampers our ability to generate large-scale datasets for computational analysis. Archives NZ series 8333 is the central filing series for the NZ Department of Internal Affairs and includes many applications for naturalization dating from the midnineteenth century onward. However, series 8333 cannot be harvested as a whole as it contains 165,352 records—well beyond Archway's limit of 10,000 results per search. Filtering the results by the keywords "naturalisation" or "naturalization" reduces the number of results to 37,674. To reduce the number of results to under 10,000, we decided to limit the search by date as well and harvest records matching "naturalisation" or "naturalization" dated between 1840 and 1905.[24] This returned 8,288 results.

An alternative approach to this problem is to slice up the complete result set using another field, such as a date or record identifier, then loop the screen scraper over these subsets and combine the results. Tim's RecordSearch harvester includes some examples of harvesting large series using this method. In the case of Archives NZ series 8333, we tried splitting the results into chunks by year, but after harvesting, combining, and deduplicating the results, we found the total number was less than expected. We do not know why. This highlights a more general problem in generating collection datasets through screen scraping. In order to code a scraper, we have to make assumptions about the way the system works based only on what is delivered through a web browser. We do not have access to the system's internal logic or data structures. This can cause problems that are only revealed when we try to do something with the harvested data, such as attempting to aggregate naturalization data by name, date, or place.

[24] The start and end dates of 1840 and 1905 were chosen as they broadly correspond to the "colonial period" of naturalization under consideration in Kate's wider project on Chinese colonial citizenship. The somewhat arbitrary selection of this set of dates highlights the necessity of working around existing archival systems. If Archway returned more than 10,000 search results, these kind of search parameters would not have been necessary.

Undocumented Transformations

The online naturalization index on the NSW State Archives website provides another example of the sorts of difficulties encountered in compiling data from collection databases and archives websites. The online index comprises information taken from an original paper index which was compiled from certificates of naturalization issued in New South Wales between 1849 and 1903. The volumes of the original index have not been digitized, but their transcription for the online version has provided a useful source of naturalization data in a structured form. However, while search results from the online index look like a dataset, with clearly defined rows and columns, they cannot actually be downloaded in a machine readable form. Nor is it obvious how to browse the complete index, rather than search for specific names; such name indexes are often compiled with family historians in mind, researchers who are searching for one or two particular individuals. Another of Tim's screen scrapers can extract data from this and other NSW State Archives indexes, and currently 60 indexes have been harvested in this way, with 1,488,222 rows of data shared in CSV formatted files.

The NSW State Archives online naturalization index is particularly interesting for analysis as it identifies the country of origin (or "native place") of applicants. After aggregating Chinese place names (such as "Canton," "Amoy," and "Hong Kong"), we were quickly able to visualize Chinese applicants for naturalization over time as a proportion of the total number of applicants. Except something was wrong. Kate noted that, based on her own manually collected research data for New South Wales, there seemed to be too many Chinese naturalizations. After further analysis we realized that many Chinese names were duplicated. A name like "Ah Gee," for example, is listed twice—once with "Ah" as the surname and "Gee" as the first name, and again with the name order reversed. Was this how names were listed in the original paper index, or was it a result of the transcription process? Either way, it reflects a profound misunderstanding of Chinese Australian names, and offers a further example of how metadata created for discovery cannot be used for other forms of research without careful interrogation.

Lisa Gitelman and others have noted that "'raw data' is an oxymoron."[25] No matter how well structured or controlled, data will

[25] Lisa Gitelman, ed., *"Raw Data" Is an Oxymoron* (Cambridge, Massachusetts and London, England: The MIT Press, 2013).

always bear the marks of its creation. As historians, we need to learn to read these marks and subject CSV files to the same sort of critical appraisal that we employ when approaching a new collection of primary sources. We need to become more attuned to the processes of inscription and transformation that create archival data. Some of these transformations, like transcription and Optical Character Recognition (OCR), mobilize the text content of records for use in new contexts. The development of crowdsourcing tools and platforms have made volunteer transcription an important means of improving the discoverability of archival collections. We have used such a platform ourselves to extract structured data from identity records in the National Archives of Australia used in the administration of the White Australia Policy.[26] But how do transcription tools and user guidelines shape the results? Is the transcribed data subjected to processes of verification or moderation? Such questions become increasingly important as the purpose of the transcribed data moves from discovery to analysis.

The NSW State Archives naturalization index has been through multiple transformations. The series notes observe that the paper index was probably first created to fulfil the Colonial Secretary's obligation, under naturalization legislation, to create "proper indices to such certificates."[27] However, the addition of notes that record when certificates were "impounded" indicates that the index was also used in control of immigration.[28] While no digitized images of the paper index are available online, microfiche copies are included in the widely available Archives Resource Kit.[29] There are no details of the transcription process on the NSW State Archives website, but it seems

[26] See *The Real Face of White Australia*, accessed March 14, 2020, https://transcribe.realfaceofwhiteaustralia.net.

[27] See "NRS-1042: Index to Registers of Certificates of Naturalization and Lists of Aliens to whom Certificates of Naturalization have been issued," State Archives and Records New South Wales, accessed March 14, 2020, https://www.records.nsw.gov.au/series/1042.

[28] The "impounded" certificates are held by the National Archives of Australia in series A806. Similarly impounded colonial naturalization certificates for Victoria, Tasmania, and South Australia are held by the National Archives of Australia in series A801, A804, and A805.

[29] The Archives Resource Kit consists of microfilm copies of NSW State Archives' most popular and heavily used colonial records, including records relating to convict arrivals, assisted immigrants, electoral rolls, naturalization, and land grants. It is held by 40 community access points, mostly public libraries, across New South Wales. For more information see "Archives Resources Kit (ARK)," State Archives and Records New South Wales, accessed March 14, 2020, https://www.records.nsw.gov.au/archives/collections-and-research/guides-and-indexes/archives-resources-kit-ark.

reasonable to assume that the index was selected for transcription based on its value to family historians. Comparing the microfiche with the transcribed version online, it becomes clear that the data underwent a significant change. The original volumes included a single column for "Name in full" and this appears to have been split into "Surname" and "FirstName" as part of the transcription process. This seems to be where the Chinese names were duplicated. The transcribed index was then made available on the web as a searchable database. The interface to this database has changed at least once, though the basic functionality seems to have remained the same. A final step in the process of transformation has been our scraping of the index data from the NSW State Archives website and publishing it as a CSV formatted file on GitHub.[30] Once we became aware of the duplicate entries, we used Pandas, a data analysis library for the Python programming language, to remove them.

Most of these transformations to the index are undocumented. There is a series note describing the original index and some useful background information in the NSW State Archives short guide, *Naturalization and Denization Records, 1834–1903*.[31] The webpage about the online index adds information on the inclusion of item numbers and the meaning of "impounded," but there is no detailed information on the dataset itself. All we are told is that there are "5,000 +" entries. Tim's screen scraping code used to extract data from the website is available in his GLAM Workbench, and preserved in GitHub, along with instructions for anyone wanting to replicate the process.[32] The Workbench lists the number of rows harvested from each index—9,860 entries were extracted from the naturalization index. The code and data for this article is also stored on GitHub and includes the index data as harvested, as well as a "cleaned" version that excludes empty columns.[33] This provides an additional check for users in case the cleaning process removed something useful. GitHub is a version control system, so it records all changes to the code and data, and by browsing the history of the GLAM Workbench repository users

[30] Tim Sherratt and Kate Bagnall, Naturalization Data Stories, GitHub repository, 2020, https://github.com/wragge/naturalization-data-stories.

[31] State Records Authority of New South Wales, *Naturalization and Denization Records, 1834–1903: Short Guide 9* (Sydney: State Records Authority of New South Wales, 2001).

[32] See Tim Sherratt, GLAM-Workbench/nsw-state-archives (Version v0.1.0), Zenodo, accessed March 14, 2020, http://doi.org/10.5281/zenodo.3549129; Sherratt and Bagnall, Naturalization Data Stories, GitHub repository, 2020, https://github.com/wragge/naturalization-data-stories.

[33] Sherratt and Bagnall, Naturalization Data Stories, GitHub repository, 2020, https://github.com/wragge/naturalization-data-stories.

can see that the indexes were originally harvested by Tim in 2016, with updates in 2018, and 2019. All previous versions of the scraping code and data files are publicly accessible. The provenance of this dataset is complex and fragmented, and what we do with it is part of the story.

The data that we work with as historians has its own history which affects how we can access and use it. Both the Queensland State Archives (QSA) and Libraries Tasmania provide machine readable versions of their own naturalization indexes. In both cases, the files are available for download, not from their own websites but from government open data portals. This indicates that the institutions expect that the data might be of use and interest beyond their own discovery interfaces. But again, there is little information about the structure and content of the indexes. Queensland State Archives notes that their index has been "generated from records created by the Supreme Court, Brisbane, Rockhampton, and Townsville districts as well as the Colonial Secretary's Office and the Government Residents Office."[34] When you examine the contents, it is clear that there are multiple entries for some naturalization applications, but why? Each entry includes an "Item ID" that identifies the source record within the collection database, and using this field we were able to find and aggregate data about the series used in the compilation of the index by screen scraping the details from the collection database. We now know that the index draws data from 10 different record series, with most entries coming from QSA series 5177 and 5741. These series are related—series 5177 is a register of the "oaths of allegiance" in series 5741. This seems to explain some of the duplication, but if we chart the distribution of these two series over time we see that there are significant gaps in series 5741. Do these records not exist, or have they not been transcribed? And while there are 10,344 entries from this series in the index, there are only 42 items described in the QSA collection database. Our investigation of the Queensland index and a new version of the dataset that includes the harvested series details is available in the GitHub repository for this article as well as the GLAM Workbench.[35]

[34] See "Naturalisations 1851 to 1904," Open Data Portal, Queensland Government, accessed March 14, 2020, https://data.qld.gov.au/dataset/91970fa7-d3c3-4171-a89d-410481cb90e9.

[35] Sherratt and Bagnall, Naturalization Data Stories, GitHub repository, 2020, https://github.com/wragge/naturalization-data-stories, and Tim Sherratt, GLAM-Workbench/queensland-state-archives (Version v0.1.0), Zenodo, accessed March 14, 2020, http://doi.org/10.5281/zenodo.3549574.

Missing Persons

Gaps in provenance hinder use of some of the data sources we have examined, but Library and Archives Canada's naturalization index presents a different set of challenges. The story of the digitization and transcription of twentieth-century naturalization lists by the Jewish Genealogical Societies of Montreal and Ottawa is told in some detail on the LAC website.[36] The resulting database is described as "one of the few Canadian genealogical resources specifically designed to benefit those researchers with roots outside of the British Commonwealth," and it contains 491,849 searchable name references from 1915 to 1946. What makes the index particularly useful for our study is the ability to search by country of origin of each applicant. However, there is no way to browse the names of the countries cited, so you just have to hope they have been used consistently. In the NSW naturalization index data, we found 22 different ways in which applicants' Chinese origins had been described.

No machine readable version of the LAC naturalization index is available. We made a partially successfully attempt to screen scrape data from the index, but our efforts were hampered by the index's size and its rudimentary search interface. Only the first 2,000 results from any search are accessible, and these seem to be ordered by item number. While the instructions indicate that wildcard searches are possible, in fact all queries are treated as substrings, matching any position in a field. This makes it almost impossible to break a large search into smaller sections for harvesting. Fortunately, a search for "China" yields only 482 results, so we were able to extract these records and compile a CSV formatted dataset that includes a link to a digitized version of the relevant page, delivered as a PDF file.

In exploring this data, however, we realized that some people were missing. When a family group was naturalized, the original lists only recorded the country of origin of the husband and/or father; for wives and children the field remained blank. This practice was followed in the transcribed index, without any attempt to link family records. As a result, a search by country excludes most women and children. After a number of experiments, we developed a method of adding the records of family members to our original harvest. However, because of the limits on the number of results, lack of relevance ranking, and substring

[36] "Naturalization Records, 1915–1951," Library Archives Canada, accessed March 14, 2020, https://www.bac-lac.gc.ca/eng/discover/immigration/citizenship-naturalization-records/naturalized-records-1915-1951/Pages/introduction.aspx.

matching, we simply cannot know if we found them all. To the original harvest of 482, we added an additional 144 names of women and children. We have shared the method we used to harvest and augment the data, as well as a CSV formatted version of the final dataset, in the GitHub repository for this article.[37] The replication of historical data in online archival datasets, as in this instance, can perpetuate gendered aspects of law and administration, whereby women and children can remain hidden within the archive. Archival data exists within a shifting set of contexts; as a representation of past activities it is no more neutral than any other historical source.

Uncertain Contexts

Archival data undergoes many transformations, but it also leads a series of parallel lives that connect to their original contexts in uncertain ways. As noted, the Queensland and Tasmanian naturalization indexes are published through government data portals. The link from the Queensland dataset to more information about the records is currently broken, and lands on the QSA home page. More confusingly, perhaps, is that the owner of the dataset is listed as the "Department of Housing and Public Works." The Tasmanian dataset seems to have up-to-date metadata, with a working link to naturalization records within the Tasmanian Names Index hosted by Libraries Tasmania.[38] But it is difficult to find information on the origins of the data.

In these two examples, the indexes remain within government systems, but in other instances naturalization data has traveled further afield. As noted, the fact that searchable name indexes exist for many of these naturalization records is a reflection of their value to family historians tracing non-British immigrant ancestors. This also makes them attractive to commercial providers of genealogical services like Ancestry.com. Indeed, transformation of the NSW naturalization data has continued behind Ancestry's paywall. Rather than using the transcribed index from NSW State Archives, it seems that Ancestry.com has compiled its own dataset from the original naturalization

[37] Sherratt and Bagnall, Naturalization Data Stories, GitHub repository, 2020, https://github.com/wragge/naturalization-data-stories. See also Tim Sherratt, GLAM-Workbench/library-archives-canada (Version v0.1.0), Zenodo, accessed March 14, 2020, http://doi.org/10.5281/zenodo.3549621.

[38] See Tasmanian Names Index, Libraries Tasmania, accessed March 14, 2020, https://librariestas.ent.sirsidynix.net.au/client/en_AU/names/?.

certificates. Their index contains 9,305 entries, compared to 9,860 in the NSW State Archives index, and 9,097 in our deduplicated dataset. Unlike NSW State Archives, however, Ancestry.com also provides access to digitized copies of naturalization certificates, taken from microfilm copies created by and available from the archives. We cannot know how Ancestry.com's index relates to the freely available version, because there is no direct access to their data—not only are there no machine readable versions of the Ancestry.com index available for download, screen scraping is prohibited by their terms of service. Services like Ancestry.com and the Church of Jesus Christ of Latter-day Saints' FamilySearch website monitor their web traffic for evidence of automated data capture and suspend infringing accounts. While they understandably want to protect their investments and competitive advantage, it means that their services are of little use to historians who want to analyze the data as a whole. Users are restricted to a narrow, manufactured view afforded by the search box.

Perhaps this would not matter if there were a clear delineation between what these proprietary services offer and other sources of archival data. We would at least have a better understanding of what we do not have access to. Instead we have a mishmash of search interfaces, datasets, and digitized resources, without any real way of predicting just what will be where. Archives NZ provides a digitized version of its "Register of Persons Naturalised in New Zealand before 1949," from series 8376, but it is not searchable. Ancestry.com displays no images of these records but provides a fully searchable index of New Zealand naturalizations from 1843 to 1981. Queensland State Archives serves the 26,769 entries in its naturalization index through a fairly basic search function, while Ancestry.com offers up what it claims is an unchanged version of the Queensland dataset through a more advanced interface that displays only 12,190 records. Digitized naturalization records from BC Archives for Victoria and Cranbrook are available on FamilySearch but not through their own website. However, the file lists, which provide a name index to the series, can be downloaded as machine readable data from the BC Archives, while being only available as images from FamilySearch.

Proprietary services like Ancestry.com and FamilySearch are, however, only adding another layer of complexity on top of systems that are already confusing and inconsistent. Why, we might ask, are naturalization records from the colonies of South Australia and Victoria held within the National Archives of Australia, while the records of the other colonies are in state archives? Why are some collections described at item level, while others are not, even within

the same organization? Why are some groups of records completely digitized, while others are selectively digitized, or not digitized at all?

Conclusion

Much of the discussion around the impact of digital technologies on the use of archives has related to improved discovery and access. But when we move beyond discovery to consider the possibilities of using and analyzing various forms of archival data we face a different set of challenges. As our examples illustrate, there is much work to be done in simply understanding what the data is and where it came from. Ease of discovery does not equate to ease of use. Rik Hoekstra and Marijn Koolen note that "the default perception" among scholars "is that the 'real research' happens after digital data has been cleaned, normalised, and organised."[39] As a result, the processes through which we come to understand data, to document its transformations, to know its stories, are rarely described. We believe, however, that such data stories are critical if we are to understand both the possibilities and pitfalls of research in digital collections. They are the missing links that enable researchers to move beyond curated collections of digitized resources and engage with uncertainties of archival systems.

We have explored these questions in the context of research, but they point to a larger gap in the teaching of history at undergraduate and postgraduate level. We expect students to critically interrogate primary sources, to examine their context, and the circumstances of their creation. As this article has noted, digitization and the construction of online discovery systems add new complexities to ideas of context. As historians we need to interrogate not just the source itself, but the means by which it is delivered to our browser. This is not simply a matter of improving "digital literacy" or developing students' digital skills. It is a recognition that the digital systems which underpin much historical work are themselves constructed—they too need to be *read*.

What is access? As students start to make use of digital collections they should be encouraged to ask why their search revealed some documents but not others. How does the delivery of a document—as an image, as a PDF, as text, in isolation or as part of a collection—change

[39] Rik Hoekstra and Marijn Koolen, "Data Scopes for Digital History Research," *Historical Methods: A Journal of Quantitative and Interdisciplinary History* (November 14, 2018): 1, doi:10.1080/01615440.2018.1484676.

how they see it, or how they use it? Messy or inconsistent data is difficult to work with, but what sorts of assumptions are made when data is "cleaned" or normalized? Every CSV file is an argument—students should be invited to think about how column headings, or standardized vocabularies, construct a particular perspective on the past.

In the article "Hacking Heritage: Understanding the Limits of Online Access,"[40] Tim describes a series of experiments with online discovery systems, designed to expose their limits and assumptions. Some of these experiments can be easily repeated without specific digital skills—it is a matter of knowing what questions to ask, and how to frame them using the technologies at hand. In a similar way, this article documents our journey through multiple systems, formats, and files, not to catalogue the frustrations of digital research, but to provide examples of the sorts of questions historians need to ask of digital sources.

We started this study with the aim of assembling a series of data sets from different archival collections that might enable us to investigate changing policies and practices in regard to naturalization in the Pacific Rim settler colonies over the nineteenth and early twentieth centuries. Working within an imperial legal framework, within bureaucratic systems with striking similarities, and with similar sorts of records—applications, correspondence, oaths, certificates, and registers—we hoped to be able to observe change over time. To some extent we have succeeded, and the code, datasets, analyses, and visualizations we created are all freely available in a GitHub repository for others to use and explore.[41] We have been able to visualize most of the datasets as time series, but the process has made us more aware of the contingencies and complexities of the data. Just what are we comparing? Data-based research is often portrayed as a reductive process. Metaphors such as "mining" and "crunching" suggest that data is physically molded into shape. But the sorts of investigations documented in this article are experimental and exploratory. More than anything, perhaps, counting records and making charts has helped us understand the limits and gaps in the available data and in our understanding of it.

[40] Tim Sherratt, "Hacking Heritage: Understanding the Limits of Online Access," in *The Routledge International Handbook of New Digital Practices in Galleries, Libraries, Archives, Museums and Heritage Sites*, ed. H. Lewi et al. (Routledge, 2019).

[41] Sherratt and Bagnall, Naturalization Data Stories, GitHub repository, 2020, https://github.com/wragge/naturalization-data-stories.

Kate Bagnall is a historian and Senior Lecturer in Humanities at the University of Tasmania. From 2016 to 2019, Kate was an ARC DECRA Research Fellow at the University of Wollongong, where she undertook a comparative study of Chinese colonial citizenship in Australia, Canada, and New Zealand. She has published on various aspects of Chinese Australian history and is coeditor, with Sophie Couchman, of Chinese Australians: Politics, Engagement and Resistance *(Brill, 2015).*

Tim Sherratt is a historian and hacker who researches the possibilities and politics of digital cultural collections. Tim has worked across the cultural heritage sector and has been developing online resources relating to libraries, archives, museums, and history since 1993. His creations include useful things like the GLAM Workbench, strange things like the Vintage Face Depot, and important things like The Real Face of White Australia. You can find him at timsherratt.org or as @wragge on Twitter.