Big Data and Australian history

Big History is a term that has particular resonance for historians of Australia—a continent with a 60,000-year record of human occupation and a geological history that extends a further 3,070 million years.[1] Recently historians have also begun to engage with the concept of Big Data. It is not surprising that these two terms are often linked. Any attempt to unite natural and human history in 'a single, grand and intelligible narrative' will necessarily result in the engagement with a lot of data.[2] While few historians have access to sources of information that are so large and complex that they defy traditional means of processing and handling, much research that engages with what might genuinely be described as Big Data has an historical dimension.[3] Climate science, analysis of criminal justice statistics and life course and intergenerational health research are all good examples. This special forum in *Australian Historical Studies* on *Big Data* is thus most timely. It explores some of the ways that the increased availability of digital data is impacting on Australian historical research and focusses on digital research that connects Australia's history to wider international and transnational developments.

Big Data as a phenomenon has a history of its own. The time it has taken to double Europe's stock of stored information has decreased from 50 years in the decades following the invention of the Guttenberg press to currently just three.[4] Some have argued that the recent explosion in the generation of digital data is not the first Big Data revolution. There was a significant acceleration in the collection, production and analysis of information in the early nineteenth century—a phenomenon that coincided with the establishment of colonial settlement in Australia.[5] Indeed, a case could be made that Australian convict records were at the forefront of that revolution—they were certainly amongst the first Anglophone record keeping systems to use unique identifiers in an attempt to track individuals over their life course.

[1] M. Chazan, *World Prehistory and Archaeology* (Abingdon: Routledge, 2016):147-8 and M. van Krankendonk, R. Smithies and V. Bennett, *Earth's Oldest Rocks: Developments in Precambrian Geology* (Amsterdam: Elsevier, 2007): 91.

[2] W.H. McNeill 'foreword' in D. Christian, *Maps of Time: An Introduction to Big History* (Berkeley: University of California Press, 2011): iv.

[3] C. Schöch, 'Big? Smart? Clean? Messy? Data in the Humanities', *Journal of Digital Humanities*, 2, 3 (2013): http://journalofdigitalhumanities.org/2-3/ [accessed 10, 6, 2016].

[4] V. Mayer-Schönberger and K. Cukie, *Big Data: A Revolution that Will Transform How We Live, Work, and Think* (London: John Murray, 2013): 10

[5] I. Hacking, *The Taming of Chance* (Cambridge, Cambridge University Press, 2002):2-3.

The speed with which metropolitan data collection initiatives spread to Britain's colonies is notable. This included attempts to conduct regular musters and censuses as well as the compulsory registration of births, deaths and marriages (first introduced in Van Diemen's Land in 1838, one year after England and Wales). From 1824 all British colonies were required to submit annual returns to the parliamentary blue books, a trans-imperial system of data collection that in the period to 1950 generated something in the order of 1.5 million pages of statistical information. Colonial attempts to emulate metropolitan practices were often extended to a greater proportion of the population. This was certainly the case with the *Police Gazettes* introduced in every Australian colony in the second half of the nineteenth century. Modelled on the British *Hue and Cry* these publications carried descriptions of convicted and discharged prisoners on a scale that could not be contemplated in the more densely populated British Isles.

Although nineteenth-century states attempted to track population wide trends in mortality, housing density, age structure and occupation, some groups of individuals were subjected to particular scrutiny. This was especially the case with slaves, indentured servants, prisoners and soldiers, and it is no coincidence that the records generated by these institutions are central to the contributions that form this special edition.

Yet, despite the desire to use record keeping systems to control subject populations, the numerous returns that flooded into administrative offices each year out-stripped the ability of states to digest the information collected.[6] Such access problems have significantly diminished in recent times. Historical records are being converted into machine readable form at an impressive rate. It is estimated that within a decade most books printed before 1900 will have been digitised.[7] The huge expansion in family history has created incentives for commercial companies to digitise large volumes of records, including births, deaths and marriages, census data, criminal records, street directories, passenger lists and and other record group that name large numbers of people. The amount of digitised newspaper content also increases each year as a result of initiatives like the acclaimed National Library of Australia, Trove project. Much of this data is live in the sense that users contribute to it—either by correcting information or adding new content. Records for over 700 million individuals have been linked and organised into

[6] T. Richards, *The Imperial Archive: Knowledge and the Fantasy of Empire* (Verso, London, 1993): 3-7.

[7] T. Hitchcock, 'Confronting the Digital', *Cultural and Social History*, 10, 1 (2013): 9.

family trees by users of Ancestry.com for example—a trend that shows little signs of abating.[8] Yet, while digitisation has made the products of a Nineteenth-Century data revolution more accessible, that access has come at a cost.

Although some historians have been able to work with commercial data providers, others have found themselves locked out of the digital archive by paywalls and search structures which are unhelpful for the kind of analysis that they wish to conduct. Most commercial data providers will only invest in the digitisation of content that will help family historians locate potential ancestors. The datasets they produce can be searched by name, but not other variables like occupation, crime and height. There are also problems with the quality of digital data. Optical Character Recognition software can produce particularly problematic results when applied to newspapers. It is not uncommon for more than 50 percent of digitised words to contain at least one error, although software developments and the use of crowdsourcing are likely to lead to improvements over time.[9]

One of the lessons of the Big Data revolution is that a lot of what is generated is decidedly messy—although this may not be immediately apparent to the end user. A particular issue is the manner in which digitisation removes items from their archival context. This can encourage researchers to treat every piece of retrieved information with equal importance.[10] The aggregation of data from multiple sources can lead to important analytical breakthroughs, but can also magnify the problems associated with the uninformed mining of archival content in the absence of context. In other words, a failure to understand the political, cultural and social assumptions (and constraints) that shaped the creation and historical use of records will blunt the power of any analytical exercise. Access to ever increasing amounts of digital data will certainly open up research opportunities, but the effective utilisation of that information will rely more than ever on the ability of end users to historicise the data at their disposal.[11]

[8] D. Gou, A.B. Kasakoff, C. Koylu, Yuan Huang and J. Grieve 'Historical Population Informatics: Comparing Big Data of Family Trees and the U.S. 1880 Census for Migration Analysis', https://dmm.anu.edu.au/popinfo2015/papers/1-guo2015popinfo.pdf [accessed 10 June 2016]

[9] Hitchcock, 'Confronting the Digital': 13.

[10] A. Gailey, 'Some Big Problems with Big Data', *American Periodicals: A Journal of History and Criticism*, 26, 1 (2016): 22-4.

[11] J. Grossman, '"Big Data": An Opportunity for Historians', *Perspectives on History*, 30, 3 (March, 2012), https://www.historians.org/publications-and-directories/perspectives-on-history/march-2012/ [accessed 10 June 2016].

To this end it is helpful to be able to compare differences in the way data are distributed across record sets. The ability to align information collected by different institutions, or at different times, can be richly informative. Having access to census data provides the opportunity to place records collected to control subpopulations within a wider societal context. The information created in order to manage nineteenth-century prisoners, for example, is a particularly rich source of social data for working-class women. In order to draw wider inference from these records, however, it is important to first establish how representative the individuals whose lives were captured in these sources were of the population as a whole. As Lucy Williams and Barry Godfrey demonstrate in their article on the UK censuses, access to complete count census data can be particularly important in enriching and contextualising these and other sources of information.

Since record keeping systems functioned as a tool of colonial, as well as metropolitan control, the digitisation of archival resources has provided increasing opportunities to explore a range of critical trans-national questions. Several contributors to this issue explore the use of datasets that are large, not just in terms of record numbers, but also their geographical scope. For instance, Catherine Hall's article uses the British slave compensation database to explore the trans-imperial legacies of slavery is a case in point. Likewise, Kris Inwood and Andrew Ross' analysis of First World War enlistment records reveals the ways that such data can be used to drive comparative assessments of the health of migrants and colonial populations. Digital history is playing a powerful role in linking Australia's past to a wider historical and imperial agenda.

The extraction of data from multiple archives can be a laborious but fruitful exercise. As Clare Anderson's shows in her pioneering reconstitution of convict forced migration in the British Empire in the period 1787-1839, wider inferences about the relationship between penal transportation, slavery and indenture can be drawn. Mark Finnane and Andy Kaladelfos also use long runs of data to place contemporary issues within an historical context. Utilising 125 years of court records and newspaper reports of criminal justice proceedings, they reconstruct the history of the prosecution of indigenous defendants charged with murder in Western Australia. Other contributions examine the potential of historical data to explore life course and intergenerational socio-economic outcomes at the individual level.

There is no doubt that digitisation has assisted the process of systematically collecting, coding and linking data. Australian research provides a good example of the increase in magnitude of research datasets assembled by researchers over time. In the 1960s Lloyd Robson assembled a sample of 7,379 convicts transported to New South Wales and Van Diemen's Land which he analysed using punch cards. Given the limits of the available technology, this was an heroic endeavour.[12] In the mid 1980s the *Convict Workers* team at the University of New South Wales digitised a sample of 19,711 convict indents—a dataset large enough to put strains on the University computing service who were thanked for making 'special concessions to one of their larger users.'[13] By way of contrast, there are currently 87,470 arrival records for convicts and assisted migrants in the *Founders and Survivors* database and the project as a whole holds in excess 1.3 million linked records pertaining to individuals who lived in Tasmania in the years to 1924. Despite this increase in scale, the data analysis for this project can be conducted on laptops. While the digital transcripts analysed by the project team do not presented a storage challenge, the many thousands of images of original records do. If *Founders and Survivors* represents a departure from past quantitative explorations of convict transportation, it is through attempts to ensure that electronic transcripts and coding systems are linked to digitised images of the content from which they were derived. Some of the challenges involved with this have been overcome through the close partnership the project has developed with the Tasmanian Archives and Heritage Office. A significant benefit of this is that it has resulted in ever improving public access to a UNESCO Memory of the World registered archive.

It is no accident that all of the articles that make up this issue have their genesis in wider and often international collaborative initiatives. These include: *Legacies of British Slave-ownership,* University College London; *The Carceral Archipelago,* Leicester University; *The Digital Panopticon,* Universities of Liverpool, Sheffield, Oxford, Sussex and Tasmania; the *Prosecution Project,* Griffith University as well as *Founders and Survivors,* universities of Tasmania, Melbourne, Monash, Guelph and Liverpool. The digital data revolution is likely to bring disproportionate rewards to those who operate within teams. While the collection of a lot of data can be an expensive undertaking, economies of scale can be achieved by those who dare to think big. A team working to a common goal is likely to be more efficient than several uncoordinated individual projects. A collaborative approach is also more likely to result in the development of common coding structures and systematic rules for handling problematic cases,

[12] L.L. Robson, *The Convict Settlers of Australia* (Melbourne: Melbourne University Press, 1965): 176-213.

[13] S. Nicholas (ed) *Convict Workers: Reinterpreting Australia's Past* (Cambridge: Cambridge University Press): x.

enabling others to make effective use of the data after the life of the project for which they were originally created.[14]

It is not always necessary to capture a lot information, however, in order to take advantage of the increased availability of digitised historical material. Text mining tools that use semantic searches to retrieve content from digital archives can enable historians to efficiently chart changes in language and terminology overtime without becoming mired in "Big Data".[15] Access to large runs of digitised newspapers, professional journals, books, court records and other publications have created new opportunities for historians working in a wide range of fields. It is now possible to reconstruct individual life courses by sifting through large amounts of data on line. Catherine Hall, for example, uses this form of approach to tease out the links between assets distributed as a result of slave compensation pay outs, the migration of imperial families, and settler colonialism in Australia—a reminder that prosopography is an area of historical inquiry that has been profoundly shaped by availability of increasing amounts of digitised data. It is no longer necessary to be trained in quantitative techniques in order to take advantage of the digitisation of large amounts of historical data.

The use of GIS (Geographical Information Systems) to store, manipulate and analyse historical data is becoming more common—a subject worthy of a special edition in its own right. Historians are using 3D models to explore past interactions with the built environment. The way history is disseminated, visualised, and consumed, is likely to change in the wake of such work. Data visualisation provides opportunities for researchers to employ new analytical techniques to understand the past, but it also present opportunities to communicate their findings in non-traditional ways. The *Dictionary of Sydney*, the *Australian Dictionary of Biography*, *Digital Harlem* and the *Tasmanian Name Index* all offer examples of the way that research conducted in Australia is engaging an online public.[16]

[14] R.H. Steckel, 'Big Social Science History', *Social Science History*, 31, 1 (2007): 13

[15] See for example, E. Toon, C. Timmermann and M. Worboys, 'Digitisation, Big Data, and the Future of the Medical Humanities Text-Mining and the History of Medicine: Big Data, Big Questions?', *Medical History*, 60, 2 (2016): 294-300.

[16] The Dictionary of Sydney, http://home.dictionaryofsydney.org; The Australian Dictionary of Biography, http://adb.anu.edu.au; Digital Harlem, http://digitalharlem.org; The Tasmanian Name Index, https://linctas.ent.sirsidynix.net.au/client/en_AU/names/

While it might be argued that the academic worth of this work has been undervalued, the impact agenda is likely to bring increasing rewards to those who use digital technologies to develop industry and community partnerships. The advent of Big Data will also have implications for the way that Australian history is taught. As all of this suggests, the digital revolution will continue to change the way we work as historians. Fortunately, the principles that lie at the core of good historical research should ensure that the discipline is in a good position to deal with the challenges that will inevitably accompany the increasing availability of digitised data, as well as enabling the discipline to capitalise on the associated exciting benefits.

Hamish Maxwell-Stewart
University of Tasmania