Building a Life Course Dataset from Australian Convict Records Founders & Survivors: Australian Life Courses in Historical Context, 1803-1920

Janet McCalman, Centre for Health & Society, University of Melbourne

Rebecca Kippen, Centre for Health & Society, University of Melbourne

Leonard Smith, Demographic and Social Research Institute, Australian National University

Sandra Silcot, Centre for Health & Society, University of Melbourne

Abstract

Founders & Survivors is a multi-university and public collaborative project that is building a transnational and inter-generational dataset of life courses generated from the UNESCO recognized convict records of Tasmania. This paper outlines the technical history of the project: mass digitization and archiving online of over 100,000 images; manual scholarly transcription; TEI standard XML data library based on automated and manual record matching and linkage; crowdsourcing using Google Docs to manage over fifty-four online volunteer genealogists; re-constitution of amalgamated life courses and record linkage; development of customized genealogical database for population and family analysis (Yggrasil); export to statistical programs (SPSS, Stata). Manual linkage and scholarly verification remained essential for the collation of prosopographical data and manual coding, derived from historical analysis was necessary for statistical analysis.

Background

The Australian colonies were settled using the forced labour of convicts. Between 1803 and 1853 almost 73,000 men, women and children were transported to an island prison on Van Diemen's Land, now the Australian state of Tasmania. The convicts' discipline, health and labour were managed by a 'paper panopticon' and those convict records are recognized in UNESCO's Memory of the World.¹ The Tasmanian records, in particular, are arguably the most detailed and intimate records of human beings' bodies and behaviour in the nineteenth-century world. Since 2008, the project Founders & Survivors: Australian Life

Courses in Historical Context has been building a prosopographical life course dataset of that population captured by transportation, starting from the surviving records: an archive that comprises around 68,000 individuals.(Bradley 2010)² Such an endeavour was impossible before the Internet and the explosion of digitized historical records made available by both government agencies and commercial genealogy enterprises. Information technology has enabled research on a scale that was unimaginable twenty years ago, yet the project has still needed hundreds of thousands of human hours to decipher, transcribe, interpret, code and connect data: machines could help us create the data libraries, but not in the end, do the research. This paper outlines the design and conduct of the project and links to web documents on the information technology.

The Tasmanian Convict Records

The penal system needed precise accounts of convicts' bodies and faces: colouring, shape, the size of facial features, and of every bodily scar or deformity, tattoo or mark to identify convicts in the absence of photographs and fingerprints. In the creation of the 'indent' or record of 'landed human cargo', the convicts provided testimony of their convictions and confessions, of their birthplaces, their families, of their religion, and their past associations. Penal transportation was forced labour so that convicts' work experience and potential usefulness to the colonial economy were recorded, records of their work assignments were created and they were subjected to a muster or census periodically. Finally, there was the conduct record that tracked their

¹ <http://www.unesco.org/new/en/communication-and-information/flagship-project-activities/memory-of-the-world/register/full-list-of-registered-heritage/registered-heritage-page-8/the-convict-records-of-australia/#c186408>

² Bradley, J., Rebecca Kippen, Hamish Maxwell-Stewart, Janet McCalman, Sandra Silcot (2010). "Research note: The founders and survivors project." *History of the Family* 15(4): 467-477.

'moral career' through servitude, recording infractions against convict discipline, secondary crimes and punishment. These are complex documents, created from the reports sent in weekly to the Convict Department in Hobart by the police magistrates who administered the system of justice.³

The Tasmanian records fall into two systemic categories: the **Assignment** period (to-1840 for men, 1844 for women) where the management of convict labour was outsourced to private employers, and the **Probation** period (1841-1853 men and 1844 -1853 women) where convicts served a term of probation under penal supervision before being assigned to the labour market. The records for these periods differ in their detail and organization. There are also musters (the convict censuses), records of permission to marry while still under the system, and inquests. Other records of tickets of leave, pardons, absconders, secondary offences and complaints were published in the police gazettes of both Tasmania and Victoria and were also published in the newspapers. In addition to the convict department's records created in the records were created before embarkation and are mostly held in the National Archives of the United Kingdom. These include some convicts who were withdrawn from the voyage and never embarked, but there was an attempt to reconcile the records created in the place of sentencing and in the penal colony. In addition there are hulk records and prison records held in the United Kingdom and Ireland. Among the richest records are the ships' surgeons' medical logs for the voyage that record individual medical cases, individual character assessments as well as general remarks on health and behavior at sea.(Foxhall 2011)⁴

These records are handwritten and are replete with arcane terms, codes and abbreviations. They are preserved in large bound volumes on strong rag paper, with cross hatching

annotations in multi-coloured inks. Digital imaging of the records was essential to the research program and the first task of the Founders & Survivors project when it received its initial funding from the Australian Research Council, was the systematic imaging of the records, linking the images to the convict index created by the Tasmanian Archives and Heritage Office (TAHO). This index lists every document and page where the convict appears in the archives. The digital imaging and archiving were managed by technology developed by the eScholarship Research Centre at the University of Melbourne.⁵

The task of data capture was therefore complicated by the multiple sources for each convict. If a life course were to be assembled for an individual, it would need to incorporate these various documents. The next task of the project was to transcribe identifying data for each convict from descriptions lists and indents to define the population for study. Trained transcribers, mostly post-graduate history students from the University of Tasmania, did this work under the rigorous supervision of Dr Alison Alexander.⁶ They transcribed the assignment records, recording name, place of birth, place of conviction, crime, sentence, age on arrival in the colony, height, eye colour, hair colour, scars and tattoos and religion. The assignment records were created in alphabetical volumes that included multiple voyages, hence the finding of a particular convict's record depended on the TAHO index references. For the probation period starting in 1841, record keeping became more elaborate and each voyage was recorded in a single volume, with a convict per page. Finding an individual was easier and the project was fortunate that Dr Deborah Oxley who had amassed an extraordinary database in the 1990s from the microfilm of indents held in the Mitchell Library, Sydney, gave permission for it to be incorporated.⁷ There were incompatibilities between these two initial data capture projects that have caused technical problems, but the Oxley records included more data on the convicts' families as recorded in the indent. Since then, Tasmanian

³

<http://www.linc.tas.gov.au/tasmaniasheritage/popular/convicts/convictdept>

⁴ Foxhall, K. (2011). "From Convicts to Colonists: The Health of Prisoners and the Voyage to Australia, 1823-53." *Journal of Imperial & Commonwealth History* 39(1): 1-19.

⁵ <http://www.esrc.unimelb.edu.au/>

⁶ <http://alisonalexander.com.au/>

⁷ <http://www.all-souls.ox.ac.uk/people.php?personid=47>

project leader, Associate Professor Hamish Maxwell-Stewart has greatly enhanced this dataset in partnership with Dr Trudy Cowley of the Female Convicts Research Centre, adding deaths found in the convict system, permissions to marry, conduct records for 1 in 25 of the whole convict population, and records of exit from the system and later crimes. This convict dataset resides at the University of Tasmania and uses Filemaker Pro.⁸

However, the Founders & Survivors project was designed to take the life courses of convicts beyond the gaze of the paper panopticon and needed to prepare the data for genealogical and demographic research. Once the convict population's identification data were transcribed, the data had to be translated into a language that would ensure its sustainable use and flexibility. It needed to be assembled and expressed as a library of life courses that were linked to the specific records of original data. The project, under the guidance of Sandra Silcot and Dr Len Smith, decided to follow the practice of the Old Bailey project and use TEI and XML.⁹

With the unique identifiers for each record for each convict, automated data linkage produced a high level of matching for most voyages, with a manageable residue of mismatches for convicts with the same name and those whose original records were irregular or missing—a deficiency that was unavoidable in a record system that spanned almost a century and worked 10,000 miles apart and across an island that was in the process of conquest and colonization.¹⁰ The project maintained strict archival standards and transcribed the records as they were found, even if original mistakes were identified. Likewise the amalgamated database included all the transcribed data as it stood in the original sources. The researcher therefore could sight every scrap of information about that convict in its original

form. The research could proceed not just with online access, but with every piece of data extracted from the convict records linked to the image of the original source.¹¹ This enabled further analysis and coding of the original conduct record, for instance, overcoming the time-consuming and error-prone task of full transcription. Using the open source program Drupal, a workstation was mounted online so that researchers could find a convict, consult the transcriptions of the key data and submit additional information from genealogical and wider historical research and construct a systematic prosopography. For the Assignment period, the image each convict's records were linked to the summary of the records.¹²

Tracing Convicts before and after sentence

The greatest impediments to tracing convicts outside the convict system were aliases and common names, particularly among the Celts of Ireland, Scotland and Wales who had adopted patronymics. Spelling of family names was also unstable—again especially for Gaelic speakers undergoing Anglicization—and for the illiterate and semi-literate. Genealogical research into convicts requires a high level of both genealogical skill and historical imagination and knowledge. Researchers have to second-guess people's movements around the landscape, through the economy, within family structures and associations. It is an exercise in fuzzy searching, looking for misspellings, slight alterations in spelling, phonetic spelling, deceptions, lies, mistakes and telling omissions. It is also time-consuming and expensive, because outside Tasmania, vital registration certificates have to be purchased from the various state registrars of births, deaths and marriages.

Australian state and federal governments have not kept the household schedules of past censuses, so longitudinal research using nominal census data is impossible. However there is an abundance of other sources that are now accessible online. First, the Australian

⁸ <http://www.utas.edu.au/arts/people/people/hamish-maxwell-stewart>; <http://www.researchtasmania.com.au/>

⁹ http://www.arts-humanities.net/projects/electronic_old_bailey_sessions_p_ceedings_c1670_1778; <http://dev.founders-and-survivors.org/pub/taho-fas-meet-2013/index.xml>

¹⁰ <http://www.founders-and-survivors.org/tei/fas-xml-demos/fas-index-tei-BKU.xml>

¹¹ <http://foundersandsurvivors.org/pubsearch/convict/chain/ai55675>

¹² <http://www.foundersandsurvivors.org/>

colonies were early to adopt systematic vital registration: in 1838 in Van Diemen's Land, just twelve months after it began in England and Wales; in 1853 in Victoria and similarly in New South Wales. Victoria established a system of registration that became a world gold standard when William Henry Archer was able to implement the full regime initially recommended by Dr William Farr for England and Wales.¹³ NSW and South Australia followed suit, but not with the rigour of the Victorian system. Victorian birth certificates, for instance, list the mother's other children, alive and deceased, with ages. Death certificates include full details of the deceased's place of birth; parents and their occupations; time in Australia broken down into various colonies; marriages (place, age at marriage, name of spouse); and children of those marriages, alive and dead, with ages. In the case of death certificates, the quality of the information depended (and still does) on the knowledge of the witness or the family history that has been passed down as acceptable. Therefore death certificates often have to be read 'against the grain' to find convict pasts that we being rewritten within families and communities.

Resources

The Founders & Survivors project began with a database of births, deaths and marriages for Tasmania from 1845-1899 transcribed twenty years before by Dr Peter Gunn from the original documents held in the Tasmanian Archives and Heritage Office. Dr Rebecca Kippen later expanded the linkages and used that database for her doctoral dissertation on the reporting of death in nineteenth century Tasmania.¹⁴ This database provides causes of death and some identifying information for individuals—sufficient to find most convicts who died in Tasmania before 1899. Since the mid-1990s digitized pioneer indexes have been available for Tasmania, Victoria, and New South Wales, and since around 2010, *Ancestry.com* has published them

as the Australian Indexes to Births, Deaths and Marriages with a steady updating with new data. Coverage is now good for Victoria and NSW, with deaths publicly searchable until 1985, and marriages and births restricted by privacy legislation to a hundred years after the event. Tasmania is still patchy after 1899. Queensland and South Australia are improving and are better for deaths than other vital events. New Zealand has recently made historical death certificates searchable online, although *Ancestry.com* does not yet include their index. *Ancestry.com* has therefore provided quicker access than the use of multiple CDs. *FindMyPast* provides the best access to British and Irish data but both companies have database searches that enable searching censuses, births and baptisms, marriages, convictions and time in hulks. *FindMyPast* has digitized some of the Irish prison registers and its British Papers online has proved very useful, while *Ancestry.com* has digitized Australian electoral rolls. But the digitization project that has transformed this historical research is the National Library of Australia's historical newspapers' database that now covers 518 metropolitan and regional newspapers. The technology is OCR from microfilm that is accompanied by a tidy interactive program that transcribes the text and invites users to correct it, while keeping a wiki-like record of corrections and registered editors.¹⁵

Crowdsourcing

It was clear from the beginning that the task of tracing large numbers of convicts was impossible without volunteers. Funding to pay research assistants on that scale is unattainable in Australia and we estimate that the project has captured something like \$4 million AUD worth of research assistance. Volunteers were also necessary to provide the links from present day families back to convicts who had concealed their past with aliases and false trails. The tracing and verification of individuals required finding triangulating data of birth dates, birthplaces, marriages, associations and occupations. Death certificates were the most reliable sources, but the researchers were able to find newspaper

¹³ Hopper, J. (1986). "The contribution of WH Archer to Vital Statistics in the Colony of Victoria." *Australian Journal of Statistics* 28: 124-137.

¹⁴ Kippen, R. (2002). "An indispensable duty of Government: civil registration in nineteenth-century Tasmania." *Tasmanian Historical Studies* 8(1): 42-58.

¹⁵ <http://trove.nla.gov.au/newspaper/result?q=>

and other vital registration data to confirm or disprove a match. Trained data checkers undertook the final verification.

When new funding was awarded in 2011, it became feasible to extend the tracing work to build a systematic reference population by dividing the convict population into historical cohorts created by their shared experience on the convict ship, and researching every convict on that voyage: the Ships Project. This reference population would enable us to assess the significance of those found and those not traced and impute results where necessary.

Over a research period of twenty months, fifty-four volunteers from all states of Australia and even the United Kingdom completed 91 ships from 1812 to 1853, yielding research populations of 12,068 men and 5,549 women, or around 25 per cent of the women and 45 per cent of the men in the convict population with intact records. In addition the database contains another 6000 convicts from other ships, most of whom were convicts traced to a death and or a marriage, through relationships with convicts in the reference population. Others were contributed to the project by descendants or discovered by other research activities, such as *Police Gazette* entries. At the time of writing the full dataset is over 23,500 and with addition of deaths under sentence or at sea, should reach 35,000 by the end of the project: around 50 per cent of the convicts whose records have survived in the Tasmanian collection. And we will have spent more than \$100,000 AUD on purchasing death certificates when the funding is exhausted.

From prosopography to population analysis

The prosopography database is a repository of historical text in XML, now with millions of lines of data. The data as it stands has limited analytic capacity: numbers for ages, heights and length of sentence, but unstable texts for everything else, including place names. The indents were transcribed by clerks from oral testimony from the convicts, hence the spelling of place names was phonetic and interpreted through thick regional accents and vernacular pronunciation. The clerks, themselves convicts, were not conversant with most of the birthplaces of the subjects. One transcribed 'Obon' [Oban, Argyle Scotland?] as a place of

birth for a cockney girl instead of Holburn, London; 'Coxhill' was in fact Coggeshall in Essex. Irish village names entirely defeated the clerks who recorded simply the county, often not making clear whether the native place was a town or a rural parish. For conventional textual analysis, the prosopographies contained multiple records of the one convict that at times were contradictory as well as complementary, so that the data might need to be 'tidied' and collapsed into uniform language. However this would compromise the integrity of the original sources and perpetuate researchers' mistakes.

The project experimented with artificial intelligence, or cluster analysis, using a self-learning algorithm developed by a team at Monash University led by Daminda Alakahoon. This clustering tool is ingenious and very flexible, but it could not work on handwritten text, and the transcribed text proved so variable and unstable that it was scarcely cost effective. (Alahakoon 2000)¹⁶ The technique was capable of selecting cohorts with specific characteristics: e.g. Lincolnshire poachers, but its lack of precision and the fuzziness of data entered by many hands over a long historical period meant that it picked up convicts with prior convictions for poaching and missed those where poaching was described differently. The Filemaker Pro database in Tasmania can find discrete cohorts such as blue-eyed male Lincolnshire poachers, but beyond that relational databases are limited in the level of analysis they can enable. Clustering remains useful but was not very helpful for the life course analysis that was the key rationale of this longitudinal study.

Demographic and historical analysis required that the data be categorized into multiple variables, coded for characteristics and outcomes and interpreted according to research hypotheses. These research hypotheses and the coding have to be defended as part of the research process and are particular to the historical period and institutions under

¹⁶ Alahakoon, D., Halmaguge, S., Srinivasan, B. (2000). "Dynamic self-organizing maps with controlled growth for knowledge discovery." *Neural Networks, IEEE Transactions on* **11**(3): 1045-9227..

examination.(Kippen et al, forthcoming)¹⁷ Since our interest is in the life course and the effects of different life stages on survival and family formation, we have divided the convicts' life courses into (1) Life before sentence (2) Life under sentence (3) Life after sentence. Data had to be extracted from the mixture of descriptive and quantitative text collated on a convict, interpreted and coded so that one can infer life course effects.

This coding exercise was conducted by the volunteers using spreadsheets hosted by Google Docs that linked to the record system transcribed and created from the digitised images (**pubsearch**). The spreadsheets were generated from the convicts on each ship. Each convict's amalgamated record has a unique id that is hot-linked to the spreadsheet. When researchers worked on a convict, they filled out a new contributed entry (**CCC**) that recorded life events with source: births, baptisms, marriages, children's births, residences, occupations, death data, descendants who were found to have served in World War I, linked to their digitized service record in the National Archives of Australia, and finally general biographical data as free text, which most commonly are extracts from the digitized historical newspapers. Many entries are extensive and rich records of life after sentence, and standard historical sourcing has been enforced. The **CCC** therefore provides a textual history of the data found about a convict and his or her descendants, which is then hot-linked to the Google Docs spreadsheet for each ship where that data is summarized numerically and coded.

The volunteers coded the spreadsheet drawing on data from the convict records in **pub search** and from the digitized images of conduct records, descriptions and appropriation lists. This required considerable paleographic skill and a knowledge of the argot of the penal system. Rather than ask volunteers to transcribe and then code the full conduct record, it was decided that it would be more efficient to ask them to count and code clearly

visible events: appearances before a police magistrate which were dated and underlined. Events relating to those like 'stripes', days in 'solitary' or the 'cells' and offences such as 'insolence', 'drunkenness', 'refusal to work' which could be aggregated or recorded for their presence or absence. They coded for level of insults suffered and recorded offences within the system such as sexual offences and drunkenness that might point to behaviour after sentence. They summarized family formation, mobility, descendants and end of life, details and sources for which were recorded on the CCC life course. Volunteers were trained in regular workshops held in Melbourne, Hobart and Launceston, and a training manual was created using animated screen capture images on Google Docs' equivalent of PowerPoint. Images were thought to be a more effective means of communication rather than textual instructions. The spreadsheet was large, with 27 columns, variously colour coded for columns that required numbers and columns that contained codes. Coding regimes were detailed on drop down menus for each column for reference and filters applied to reduce mistakes. The spreadsheets, when cleaned, can be exported to SPSS.

Crowdsourcing and building a community partnership

The volunteers were remarkable. They were a highly self-selected group of people, few of them with formal historical training, but all of them with the instincts of good historians. They were mostly retired from work, and old enough to have been schooled in copperplate handwriting (which was phased out in Australian schools in the early 1960s). The project had built a following from newspaper, television and radio publicity because the 'story' captured the imagination of many. A full-colour newsletter was produced every four months and distributed online to recruit volunteers, disseminate instructions and report results. It also enabled the volunteers and the research team to publish substantial articles based on the research, with full illustrations. Book reviews and news of other convict projects such as the Female Convict Research Centre and the Port Arthur Historical Stire added to the feast of material reaching a

¹⁷ Kippen, Rebecca and McCalman, Janet, 'A Test of Character: A Case Study of Males Convicts Transported to Van Diemen's Land 1826-1838' in Baskerville, Peter and Kris Inwood, eds, *Lives in Transition: Longitudinal Research from Historical Sources*, forthcoming, McGill University Press.

rapidly growing national and international following.¹⁸

From the beginning, it was obvious that volunteers should feel part of the research team, share in the planning and reporting, and in the intellectual life of the project. This was sustained by whole day workshops providing lunch and talks, discussion groups and coaching sessions. Each year finished with a small conference and the project will have a final conference in 2014. Volunteers proved to have a wide interest in the historical issues, brought valuable insights and findings, and did some remarkable research into elusive lives. A number have emerged as born-history writers and we will be hoping to support some in producing their own books on their ship research. And it hardly needs saying that new friendships have been built for everyone.

An exciting late development, through the initiative of Colette McAlpine from the Female Convict History Centre is the relationships she has built with scores of local and family history societies in the United Kingdom and every day there are emails with information on convicts from county and local records. We achieved our initial dream of a chain across the '10,000 miles' between the 'Old Dart' and the Antipodes.¹⁹

Lessons learned

The hope of an open website with contributions from the public was naïve. The problems with spam overwhelmed our technical staff and the system itself, and if the project manages to find a sustainable home, email contributions that are vetted and uploaded by a volunteer research team will have to suffice. The development of the interface, the data entry forms and the spreadsheet system was iterative and we did accumulate some baggage that needs removing. It is very difficult, as with all multiple-user data entry, to maintain consistency, but we have now completed a

large cleaning and authenticating exercise and have sufficient funds to pay the best volunteers to do this work. Google Docs saved us from building an expensive workstation of our own, and has been cost effective and easy to use.

Drupal for the interface has struggled under the load and none of the software proved adequate to the task of managing family trees and relationships. This we have resolved, with the support of the Australian National Data Service, by customizing a Norwegian open source genealogical program called Yggdrasil. Our system designer, Sandra Silcot, has spent twelve months adapting and extending Yggdrasil so that it can manage and store individuals who are 'created' or 'aggregated' from the sightings in historical sources and where they can be studied as a population rather than in separate lineages. Underlying this is a program for data linkage called LNKT, developed by Dr John Bass for the Western Australia Health Department.²⁰ These software packages manipulate the data for us, but the matching and discovery, the interpretation and the coding have all needed human beings.

We extend our heartfelt thanks to our colleagues in Tasmania, Guelph, Ohio, Oxford and Flinders and to Claudine Chionh. Above all we thank our volunteers.

This project is funded by the Australian Research Council.

¹⁸

<http://wwwFOUNDERSANDSURVIVORS.org/content/chainletter-no-1-june-2009>
http://FOUNDERSANDSURVIVORS.org/sites/default/files/newsletters/FASNewsletter14_2013_large.pdf

¹⁹ <http://www.femaleconvicts.org.au/>

²⁰ <http://fasconn.blogspot.com.au/>

- Alahakoon, D., Halmaguge, S., Srinivasan, B. (2000). "Dynamic self-organizing maps with controlled growth for knowledge discovery." Neural Networks, IEEE Transactions on **11**(3): 1045-9227.
- Bradley, J., Rebecca Kippen, Hamish Maxwell-Stewart, Janet McCalman, Sandra Silcot (2010). "Research note: The founders and survivors project." History of the Family **15**(4): 467-477.
- Foxhall, K. (2011). "From Convicts to Colonists: The Health of Prisoners and the Voyage to Australia, 1823-53." Journal of Imperial & Commonwealth History **39**(1): 1-19.
- Hopper, J. (1986). "The contribution of WH Archer to Vital Statistics in the Colony of Victoria." Australian Journal of Statistics **28**: 124-137.
- Kippen, R. (2002). "An indispensable duty of Government: civil registration in nineteenth-century Tasmania." Tasmanian Historical Studies **8**(1): 42-58.