Sharing the archive: using web technologies for accessing, storing and re-using historical data

Mark Finnane, Andy Kaladelfos and Alana Piper

Introduction

The possibilities created by the digital revolution are changing the way historians and other social scientists can access, analyse and represent data. In particular, the way these digital technologies enable the creation and use of big data appears likely to change many social science research questions and their answers. As one of the social sciences, historical studies share much in these respects with some of the natural sciences – and can learn from them. This article draws attention to the methodological advantages of working with large datasets collected and curated through a collaborative approach between researchers and volunteers, commonly known as citizen scientists, but for the purposes of our research we have referred to them as 'citizen historians'. We highlight the challenges encountered along the way in creating such large datasets and how our research team has sought to address them. The article is offered as a contribution to scholarly dialogue about crowdsourcing (a term first coined in 2006 by Jeff Howe) research data using digital technologies.

We start by discussing some key characteristics of what we will call 'the history problem': namely the challenge of retrieving, storing and using data that is often hard to access. We then introduce our research, 'The Prosecution Project', a large-scale national comparative study of criminal prosecution in Australian history over 150 years. The project's ambition—to create a large, well-curated and enduring database for future research and public access—is presented as a prototype for future social science research, with emphasis on its use in history. We introduce the main features of the database by way of highlighting

the research design process, the web portal as a data collection point, the recruitment of volunteers, and the mechanisms of quality control and data storage, as well as issues in ensuring enduring legacy of the data. We discuss crowdsourcing as a research tool and highlight the potential ancillary benefits of this kind of community engagement for both the discipline and the public, as well as addressing questions of resourcing and infrastructure that such projects demand.

History's data problem

Historical knowledge is data intense, whether it is studied qualitatively or quantitatively. While the primary mode of historical analysis might be regarded as narrative, with an emphasis on exemplary stories that may be seen as metaphors for more general social practices, for many decades there has also been a strong quantitative tradition. Quantitative methods are largely determined by the particular research methodology: for example, modern histories focussing on the nineteenth and twentieth century frequently access large datasets, especially those associated with the kind of record-keeping encouraged by the modern state. Hence areas like economic history and demography require access to longitudinal data, tracking economic trends or the mobility of workforces, or utilising vital statistics of fertility and mortality of the kind that governments have collected systematically since the middle of the nineteenth century. Other research areas (for example, criminal justice history), are less well served by the availability of large and continuous datasets, especially of the kind that offer any more than an aggregate view of complex social processes (Godfrey et al. 2008).

In brief, there are numerous problems confronting historical projects that aspire to a social science approach to their subject matter. Government-generated official statistics vary in quality between nations and between jurisdictions within nations. These statistics provide

researchers with access to aggregated trends in crime, for example, but do not allow for deeper analysis below the categories of interest created at the time of their production. Even for those categories, there is often no sense of the relationships between them in official statistics, which might record the number of people arrested for different offences in a given year or the number of arrests that fell into different age categories, but no cross-tabulation of these or other variables. A researcher's access to case level data that would allow more complete analysis may be hindered by poor survival rates (for example hospital case records, which were rarely kept for long periods). Equally, where records survive, privacy considerations or diverse and seemingly arbitrary restrictions on access may hinder researchers. In these respects, the research barriers created by difficulties of data access are not fundamentally different from similar challenges in research domains attempting to capture contemporary data. All the same, historical data may survive in often surprising quantities and their very historicity presents unique opportunities for social science researchers – avoiding some of the practical limitations created by ethical requirements for research with living subjects. Often too, the entirety of the outcomes experienced by historical subjects can be traced in a way that they cannot for contemporary subjects whose futures are unknown.

As for any other research domain that requires access to large datasets, the quantum and longevity of historical data present formidable problems. What are we to do in the face of a major collection of administrative files, often surviving only in original handwritten manuscript volumes, relating to education, health, criminal prosecution, immigration and so on? Addressing such a challenge, however, presents the researcher with great opportunities. Even where seemingly reliable statistical records are available they may relate to only part of the dataset that is potentially accessible through returning to the original data. The possibility of triangulation of data—linking with related data and contextual sources—may advance the

research field in ways previously unimaginable. Where data access cannot be adequately or efficiently managed through digitisation or some form of computation the challenge of dealing with large administrative file systems may limit the researcher to quite small samples. In well-designed studies these may be very adequate to the research questions asked. But for longitudinal studies and for research that may require very large populations in order to produce a small number of cases of interest, the desirability of creating large datasets from original sources is obvious. That is certainly the case with research that focuses on discerning patterns and trends over long periods of time in areas like family formation, literacy, health and crime.

The response of The Prosecution Project to history's 'data problem' has been to design and build a sustainable and reusable relational database with the capacity and flexibility to accommodate a variety of source data from long periods of time. Our brief description here contextualises the later more detailed discussion of the project's approach to data collection and curation.

The project aims to investigate the criminal trial in all Australian jurisdictions (six states and one territory) over periods as long as 150 years, from the 1820s to the 1960s. The primary sources of data are registers of court appearances, typically including at the very least the name of a defendant, the offence charged, place and date of trial, verdict and sentence. Depending on the jurisdiction, some registers provide a great deal more information, including the names of magistrates and judges, the place of committal which may be a proxy for the location of an offence, the names of witnesses, and the legal defence arrangements including names of lawyers. For the relational database, each of these individual bits of information constituting the record of an appearance (a case, a trial event) becomes an individual data point located in a data table. Attributes in the data table are defined by the research team for collection from the research sources, as well as by reference

to related sources that enable enrichment of the case records. Cases are entered at the person level rather than by reference to a particular crime event. So any criminal prosecution involving a number of offenders as co-accused is entered for each person involved. Conversely for any person facing more than one charge the case remains at the personal identification level, with the database enabling entry on any individual record of multiple charges and outcomes (e.g. different sentences for particular convictions).

The definition of attributes for each data table takes account of the research questions driving the project, which may change over time. The original data source may include no more than 10 or so data points relating to each case. As noted above, this data relates to criminal procedure and outcome including the defendant's name, offence, plea, verdict, sentence, judge hearing the case, and date. Additional research (e.g. through other archival sources, newspaper reports, law reports, or police gazettes) enriches the basic case record, enabling the researchers to develop a complex understanding of the context of prosecution and its outcomes. Additional data often relates to case characteristic information about the defendant, the victim, and the circumstances of the offence. This might include sex and age information, birthplace, ethnicity, relationship between victim and defendant, the location of crime, duration of event, the use of weapon, goods stolen, and previous criminal history. The database is thus constantly growing, both in terms of the numbers of cases entered, and the range of attributes that refer to such cases. The recent addition of a query tool also enables researchers to link records of multiple prosecutions that relate to the same person across a number of years and even in different jurisdictions. Such an innovation is clearly vital to the capacity of the Prosecution Project to support research into criminological or legal historical concerns such as recidivism or the life course of individuals within and outside the criminal justice system. The potential benefits of databases able to generate linkages between different

records and integrate information from different sources are equally apparent for other areas of historical research, as well as cognate fields.

The resources required to build such a database are considerable. Infrastructure provision and design are initially costly, especially for a project that requires flexibility in terms of levels of permission to access data, and recognition of privacy provisions that frequently constrain public access to historical archival data. The human capital required to develop the database is the other demanding consideration. We discuss later the role of crowd-sourced volunteers in collaboration with the research team. But in the following pages we present a more detailed outline of the principal features of the project, its web-based design for accessing and curating data, and its focus on data linkage in the increasingly digital research environment that we inhabit.

Project aims

The Prosecution Project is a large-scale, longitudinal, multi-jurisdiction research project investigating the history of criminal prosecution in Australia. It makes use of the increasing availability from public archives of digitised sources that enable researchers to reconstruct an entire social process and to do this across a very large number of cases. In place of the dependence of researchers on official statistics or very limited samples of original records, the project has developed as a research collaboration to maximise the amount of data available and to ensure its long-term retention for subsequent research uses.

In brief the Prosecution Project aims to:

1. Reconstruct historical records of criminal prosecution in the six Australian states which have primary criminal jurisdiction and to do so for a period of up to 150 years from about 1820 to about 1970.
2. Make such a database expandable by facilitating enrichment of case records through linkage to other data sources.
3. Ensure the database is curated to a high level and is secured for future use and re-use.
4. Take advantage of the digital environment to retain the capacity to verify records by permanent linkage of case meta-data to the original sources.

The research questions driving the Prosecution Project are interdisciplinary, drawing from history, law, criminology, gender studies, cultural studies, sociology and other related research domains (Finnane & Piper 2016). Criminological concerns with understanding changing patterns of crime, policing and punishment are joined with legal historical questions about the changing process of prosecution. Accessing original data enables researchers to explore questions rarely contemplated in historical studies of the criminal justice system, such as the specific roles and impact of defence lawyers (Piper & Finnane 2017a; 2017b), or the victim characteristics associated with the prosecution of offences against person or property (Finnane & Kaladelfos 2016; Piper 2018). In contrast to the focus of official recordkeeping on the criminal offender, the data accessed by the Prosecution Project enables us to contemplate a history in which all parties to the process of the trial recover their rightful historical place. These might include not only victims of crime, but witnesses, investigating police, defence lawyers and crown prosecutors, judges and magistrates. So too the potential exists to contextualise the process of prosecution by reference to a number of historical and socio-legal factors from geo-location, to time period, changing statutory law and shifting sentencing outcomes.

To enable researchers to address such a vision is of course more easily said than done. No less than in the natural sciences the work of social observation and data collection in social sciences and humanities entails a great degree of labour in data entry, facilitated where possible by tools of discovery and systems of recording and retrieval. In the past the tools have been essentially pen and paper, the processes those of systematic recording and some principles of indexing. Good research has been very dependent on the meta-data systems found in research libraries and well-managed archives. But with digital approaches taking hold of the data repositories, researchers now have open to them the possibility of a great advance in the scope of data being accessed, stored and made available. To enable this big data future to become a reality requires a degree of collaboration between researchers and other communities, professional and otherwise as we will discuss below, as well as significant initial resource investment and a commitment to recurrent support of the research infrastructure required.

Research design process

Before undertaking the proposed research on criminal justice procedure on such a large scale, the research team initially faced the challenge of getting suitable data. Previous experience in accessing historical data for quantitative analysis had highlighted the importance of planning the database design in collaboration with information technology experts (Finnane and Garton, 1992). For the Prosecution Project director (Finnane) that experience had been in the days before widespread availability of personal computers and the World Wide Web; the data was managed in a Unix environment on a mainframe computer. By the time the current project came into contemplation the advantages of research collaboration and networking beyond the bounds of the research team were becoming even more evident. At an early stage,

following discussion with the university's 'e-research' specialists, the research team determined to develop its research plans in continuing discussion with those specialists.

Initial discussions with a business analyst focused attention on the need to determine the most cost-effective solution to accessing data that would be robust in quality and duration. There were very voluminous records available in Australian archives relevant to the criminal trial, but few were digitised and at that stage (just five years ago) none were online. The research team had acquired microfilm copies of the court registers—the original records of the higher courts that list all criminal appearances and their outcomes—for one jurisdiction and subsequently had these converted into individual digital images, each image corresponding to a page of the register. All these records were in manuscript, dating from the early nineteenth century to the mid-twentieth century. It quickly became evident that accessing such hand-written data through machine technology, such as use of OCR, was impossible. Management of the process by a double-entry approach (as used in the Old Bailey Online transcription of printed proceedings: https://www.oldbaileyonline.org/) was not feasible for this project, given the nature of the data). Some consideration was given to outsourcing the indexing of the records, but significant challenges of quality control as well as the cost of doing so turned the discussion towards an alternative vision.

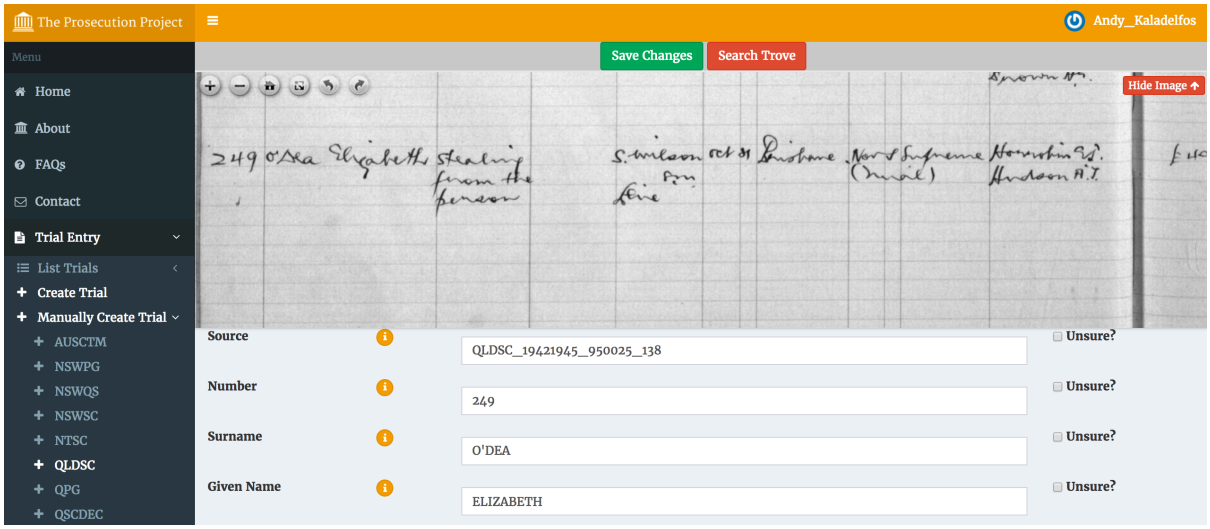

Figure 1: *The Prosecution Project Database*, https://prosecutionproject.griffith.edu.au/prosecutions (version 1, 28 August 2017), Trial ID #392628, Queensland Supreme Court, Elizabeth O'Dea, 1943.

This vision was to use the network capability of personal computers linked to the web as a means of distributing the images for indexing, initially just among the multi-member research team. The data would be entered via a web portal, with a permanent link to the source image of the data, maintained on the university's institutional server (see Fig. 1). The data would be stored in a relational database, for future access by researchers. To enable this plan to develop the research team met on a fortnightly basis, occasionally weekly, with the software engineer and web designers assigned to the project. In itself this was a major learning experience for both sides. The researchers had limited acquaintance with database design, and – having acquired further court registers from other state archives – a complex set of research data, varying in range of information across time and between places. For them it was important that the database be flexible enough to accommodate unforeseen new sources of data or variations in data availability. For the e-research specialists, the challenge was to work with a group of researchers whose approach to research design was more iterative than well-structured. Regular face-to-face meetings proved to be vital in ensuring the right degree of understanding to progress the project.

In early discussions, we explored the use of volunteers on the citizen science model to assist us in our data access work. While this was an unfamiliar model for humanities and social sciences at the time, we were aware of one very successful project that had used volunteers. This was Founders and Survivors (foundersandsurvivors.org), a database of convicts transported to Tasmania from Britain in the early nineteenth century; its work commenced in 2008. The National Library of Australia had also had considerable success in volunteer support for the transcription of OCR text from digitised newspaper images for its

Trove database (also from 2008: trove.nla.gov.au/newspaper). The research team agreed that this approach was worth exploring. In designing the database and data entry portal, the project was thus focused from the beginning on the possibility of extending the concept of 'research team' to embrace a much larger community of non-professional researchers.

Database features

The core resource of the project is the data entry site. Access to this web portal is through a secure login. Prior to login potential users are able to inspect an instructional video outlining the key features of the data transcription process. While there is no formal documentation for training transcribers, FAQs on the website provide further advice, including a glossary of terms and abbreviations common in court records, such as the Latin term *nolle prosequi* to indicate the prosecution was not proceeding with the case, or letters T.L. after a defendant's name to indicate they were a former convict who had been given a ticket-of-leave. After login, data transcription is facilitated through a clickable 'Create Trial' option that takes the transcriber directly to a web form data page, complemented by an image window that has been preloaded with a digital image of a court register. The transcriber enters data, some of which is free text, others stored in specific formats including date and URLs. Some attributes are supported with glossaries. An API enables the user to search for information in the National Library of Australia's Trove library of digitised newspapers and manually selectrelevant URL links to add to the record. Save (or 'create') commands return the data to a relational database. Only the administration team is enabled to delete records but transcribers may continue to access records they have contributed, for information relevant to the transcription of later records.

As discussed earlier, the design of the web portal facilitates expansion of the research project with a high degree of researcher control. While the original design focused on just six data tables (or registers as we name them in the project), corresponding to the six Australian

state jurisdictions with primary responsibility for criminal law, the increasing availability of alternative sources of data as well as research interest in new jurisdictions has expanded the range of the project. At time of writing there are 16 data tables, sharing a small number of common attributes such as name, offence, trial date, verdict, sentence. The web portal is the researcher interface for managing these data tables, adding new attributes to suit a particular register structure or the data needs of a new research question (e.g. details of age and gender of the accused and the victim, whether the accused was defended by a lawyer or not, the names of judge or magistrate, the number of witnesses and so on). The web portal is also the researcher interface for managing data output, conducting search queries on any of the database attributes to produce datasets for further analysis (see Fig. 2).

Showing 301-350 of 633 items.

Download | Using: [VICSC] Kaladelfos_default

Change View | Download CSV | Download TXT

Actions	Trial Id	Full Name	First Offence Listed	Date of Trial	Verdict First Offence	Sentence	Crown_Prosecutor	Defended or undefended	Defence lawyer surname
				1916					
	94669	HERBERT LEES (SCHOFIELD)	LARCENY AS SERVANT	11 Apr 1916	GUILTY	SIX MONTHS EACH 8 COUNTS CUMULATIVE FOUR INSTANCE, CONCURRENT IN FOUR INSTANCES	GURNER	UNDEFENDED	UNDEFENDED
	94652	LESLIE WINDUSS (AUGUSTINE)	LARCENY FROM PERSON	11 Apr 1916	NOT GUILTY			UNDEFENDED	UNDEFENDED
	93526	FRANK GUILFORD	HOUSEBREAKING	1 May 1916	GUILTY	12 MONTHS EACH COUNT CONCURRENT	WOINARSKI	UNDEFENDED	UNDEFENDED
	93521	ROUGHAN MCKELVEY	FORGERY	1 May 1916	GUILTY	12 MONTHS CONCURRENT	WOINARSKI	UNDEFENDED	UNDEFENDED
	93387	ETHEL ODONOHUE (MAUD)	BIGAMY	12 Apr 1916	GUILTY		WOINARSKI	UNDEFENDED	UNDEFENDED

Figure 2: Data attributes about lawyers, Victorian Supreme Court dataset.

The nature of this crowdsourced project entails a significant invetsment in infrastructure support, especially through the digitisation of data sources, their storage, their transmission to transcribers, and the retention as a permanent link to the unique case record derived from any particular image. Again the web portal has been designed with a high

degree of researcher-driven control through the 'image administration' tool. As the project has developed, new transcription capabilities have been added through a document transcription facility (see Fig. 3) – this enables the transcription of large archival files, sometimes many hundreds of pages long, including depositions and other case record materials not readily accommodated within the structured forms of the database. The web portal however also facilitates a permanent link between the primary case record and any supplementary materials that may become available through this 'trial document' transcription tool.

Figure 3: Trial Document Transcription, *R v Palin*, Western Australia 1861.

Further the web portal now also accommodates a capacity to link records between different data tables. For any one jurisdiction we already have in some cases two or three sources of potential data relating to each individual case – for example, an arrest warrant/committal charge in a police gazette, a trial record in a court register, and a discharge report including very specific biographical information drawn from prison records. A 'records linkage' tool within the web portal facilitates the generation of authoritative datasets

relating to particular individuals, enabling for example the productive analysis of co-offending, of criminal networks, or even of the shifting status of particular individuals between category of victim and accused.

In addition to the digital infrastructure, the project is viable only through significant research collaboration. This has involved not only the commitment of the research team to sharing their labour and resources productively in the building of a large database that will enable over time repeated use and re-analysis, depending on changing research questions. But also, and very importantly, the design of the project as a web-based repository of historical data has enabled the extension of the data retrieval process to include the participation of many volunteers. In this way the Prosecution Project participates in a contemporary development of research communities that embrace something more than the conventional institutional academic, working in universities and generally confined to their disciplinary boundaries and associations.

Crowdsourcing and citizen historians

The accumulation of large datasets on the basis of original observations has been tried and tested in a variety of natural sciences for some period of time. The recruitment of citizen scientists has been very successfully undertaken in astronomy and some of the environmental sciences - assisted over the last decade by the successful development of web-based platforms like the very accessible and adaptable Zooniverse (https://www.zooniverse.org), as well as the availability of transcription tools like Digivol (https://australianmuseum.net.au/digivol). The nature of the data collected has varied from the identification of unknown galaxies and other astrophysical phenomena to the transcription of old ships' logs and diaries to provide a record of climatic and environmental change over recent centuries (Showstack 2012). Mass observation of animal and bird behaviour is enhanced by the availability of data collection through handheld devices linked to reporting

protocols that prepare the data for database entry (Ellwood et al. 2015). For the most part the social sciences have come later to the possibilities opened up by these new methodologies, but crowdsourcing approaches are now expanding rapidly in the sector (Ridge 2014; Smith 2014)(Hedges and Dunn 2017).

A recent review of research conducted by crowdsourcing methods (Watson and Floridi, 2016) highlighted the growth of the methodology and the research advantages made possible by these large-scale data collections. Importantly they were able to demonstrate strong evidence of higher impact and citation rates of scientific research conducted in this way. They theorise the reason for such impact as lying within what they call the desirability of maximising evidence, or the principle of 'total evidence', meaning that analysing maximal available evidence is preferable to sampling. In place of the limitations imposed by the research efforts of individual researchers or research teams, however assiduous, mass observation enabled by well-designed crowdsourcing expands the scope of research-relevant evidence by a number of orders of magnitude.

But the benefits of crowdsourcing accumulate not only to research in general and to researchers in particular. The benefits of engagement in the work of citizen science include increased levels of science literacy, expansion of the community of science in ways that assist better understanding of the natural world and its processes, and importantly social inclusion embodied in the engagement of citizen scientists whatever their motivations. By extension these benefits might flow equally to those citizen investigators who become involved in the world of social science and humanities research. But how how might these benefits be secured? By what means this project of the kind we have described involved others outside the research team? What are their interests and motivations and how do they intersect with the objectives of the Prosecution Project?

As noted earlier, the Prosecution Project was by no means the first to engage volunteers in the work of data access for an online research project dealing with a discrete data source. An important difference from earlier examples was that the Prosecution Project was a national project covering a number of jurisdictions, with different kinds of data, recruiting online, while the Founders and Survivors project relied primarily on volunteers recruited through a local archives office as part of that institution's volunteer program. The data entry process from the beginning was also different in that the only mode of volunteers entering data for the project was to be via a web portal. Again in contrast to some other web resources accessing volunteer support, such as the National Library of Australia's very successful Trove library of digitised newspapers (Holley 2009), or the Transcribe Bentham project (Causer & Wallace 2012) the Prosecution Project at the outset could not present its data sources online to a general public. Primarily this was due to archival access conditions and permissions, as well as copyright issues. So on the one hand, the project could not plan for a general web-community of users, while on the other it had to explore the possibility of recruiting volunteers who were less likely to be concentrated around a particular institution, an archive, museum or library.

The solution for engagement of volunteers on the Prosecution Project was to develop a registration process via the project's public webpage (prosecutionproject.griffith.edu.au). A system-generated message alerts the project team to the volunteer registration, and prompts an email advice on password access to the data entry portal. Mechanisms were also developed within the system to ensure sensitive records would not be released into the general pool for volunteer transcription – these are retained for transcription by researchers approved for access.

Once the system design was achieved and tested, how was the project going to recruit volunteers? Fortunately there is a large community of interest formed around access to the

kinds of records that are central to the Prosecution Project. This is the diverse community of people interested in family and local history, many of them frequently involved in community associations focused on such interests, others engaged through commercial and other online genealogical resource services. Through a public website we hoped to capture some of this interest, but such engagement would be dependent in the first place on prospective volunteers actually knowing about the project. Consequently once our system was ready to allow volunteer participation, the project team wrote to every family and local historical society in Australia as well as the various state archives offices.

This approach was successful in recruiting a significant number of volunteers, some of them known to each other, many of them from rural and regional areas of Australia, some in quite remote regions where nevertheless their connection to the internet enables their virtual participation. As was also experienced in the crowdsourcing project Transcribe Bentham (Causer & Wallace 2012), the Prosecution Project found that while many signed up for the project, the bulk of transcription work was ultimately completed by a small number of consistent volunteers or 'super transcribers'. Many had previous experience working as volunteers with local archives; for others their express motivation (prompted by a question at registration about their reason for interest in the project) was that they had gained much from the resources available on the web about their family's or locality's history and wished to give back. Although students in related disciplines (law, history, criminology) are numbered among the volunteers, many others are retirees. This mirrors trends in the history/heritage sector more generally, with most museum volunteers in the UK now retirees rather than those seeking work experience (Holmes 2003). The project has benefited from the energy of one particular historical society, the Carnamah Historical Society (carnamah.com.au), which had already undertaken significant digital transcription on their local records, and was now interested in extending their support to other projects. Local history societies in general

appear to have been much quicker than academic historians to generate and embrace crowdsourcing projects (Grove 2010).

Retention of volunteers is a challenge, but the significant numbers signing on in the first place mitigates even a high rate of loss of interest. The result is that a process of continuous data entry has been maintained for nearly three years, with volunteers entering more than one-half of the current core data (that relating to Supreme Court records, at January 2018 more than 180,000 records). The high level of engagement of volunteers is an important signal of digital inclusion as a means of bringing researchers in closer contact with external communities. These stakeholders may have different motivations, but overlap with researchers in their joint interest in accessing new sources of data.

Although various technological aids, rewards and signals are adopted by other crowdsourcing projects to maintain volunteer commitment, the Prosecution Project has not found it necessary to go down such a path. Newsletters and social media updates for the volunteer and research community on the project's progress and outcomes have so far addressed the need for continual engagement. Particularly dedicated volunteers have also attended research seminars presented by project members; the opportunities thus created for social interaction with both each other and the research team has strengthened the sense that volunteers are not just part of a 'crowd', but a 'community' (Haythornwaite 2009)(McCalman 2013). Volunteers have also reported feeling encouraged by the growing index of records available for searching on the public website, as well as the recent addition of a tool that allows visitors to visualise statistical patterns from the data across time (https://prosecutionproject.griffith.edu.au/prosecutions/web/index.php?r=public-search%2Fvisualise); introduction of such features has also led to further recruitments. The nature of the database process, which ensures that a transcriber can always consult the records they have individually produced, together with the evidence of a growing resource for

public use via a search engine on the project website, thus appears to have provided sufficient reward to retain volunteer engagement.

The engagement of volunteers has thrown up challenges, though few that would not occur anyway in the course of the data collection and curation process. The difficulties presented by nineteenth-century handwriting have inevitably occasioned a good number of errors in transcription. While a double or triple entry methodology might address some of these, the project team has opted instead for quality assurance through regular scanning and cleaning of the data, as well as providing the greatest facilities possible for volunteers to seek help in improving their transcription work. Within the transcription page, transcribers have a self-report 'unsure' check box that can be marked against any or multiple case attributes. They have an additional opportunity to comment upon any record oddities or problems in transcribing the data at the point of submitting an entire page as completed. The project team can then provide them with advice on issues encountered for future reference; project newsletters also contain items on commonly encountered problems that the team has noted in transcription work.

A significant challenge to the authority of any dataset lies in the risk that different collection methods will affect the reliability of the data. In contrast to the standard procedure within research teams of the use of research assistants to collect and clean data, crowdsourced data entry creates a significant dependency on the prior knowledge of volunteer transcribers or their learning capacity while involved in the task. For the kind of tasks involved in the Prosecution Project experience has nevertheless shown that a good number of volunteers bring special skills to the project, including familiarity with legal forms owing to past work experience, familiarity with even quite difficult samples of cursive handwriting from up to 200 years ago (something we have found much less common among young researchers

unless they have extensive experience in archives), as well as speed and accuracy in keyboard entry.

All the same it has been important that the project remains vigilant in its attention to the quality of data. An early decision in the data entry designed was to encourage self-reported uncertainty for any category of data being entered and this has been widely used by transcribers. For some time it remained possible to moderate this self-reported 'unsure' check box and attend to the entries so identified. The rapid growth of the database nevertheless means that data correction remains a continuing burden. Ease of reporting out of the database in CSV format has enabled very efficient correction of significant repeat errors such as place names or names of judges. Batch editing of some thousands of case records remains an invaluable aid to the establishment of good quality datasets. Much of this work has also been undertaken at the point of research analysis, when members of the research team extract the dataset and then clean it in the course of preparing the data for analysis. The crucial take-away message however is that any system that relies on a wide range of human factors can ill afford to ignore a continuing attention to the quality of data.

Outcomes

What has the project achieved to date? As a research project the core aims are those relating to the research domain, in this case of criminal justice history and related concerns. But as noted ealier a core objective from the beginning was to access data and retain it in a form that could be used by other researchers in the future. So a few words may be useful to summarise the scope of the database to date and plans for its sustainability.

At time of writing (January 2018) the database now holds more than 550,000 records, from eight jurisdictions in Australia. The jurisdictions include the six states of Australia, the

Northern Territory and the Commonwealth of Australia in its military jurisdiction (for courts martial). The flexibility of the database design has also enabled the addition of register data relating to a UK jurisdiction (West Yorkshire) as part of a comparative study with Australian records. The records extend for nearly 140 years in some cases, with all but complete datasets for the Supreme Courts of three states, with the remaining states expecting completion by mid-2018. In addition the project has created or accessed datasets from a number of other related sources including police gazettes, which include records of people tried as well as those discharged from prison, prisoner registration books, and records of some lower courts dealing with criminal matters.

With the aid of the transcribers as well as the energies of the research team, many records have now been linked to other data sources, especially the invaluable historical reports of newspapers, vital for recording court proceedings in the period before court trials began to be transcribed in the mid-twentieth century. The continuity of the data as well as recent developments to the online site enable researchers and public users to visualise major trends in prosecution and sentencing over long periods of time. Future development of the data is likely to include geocoding of crime or offence locations as well as places of prosecution.

In the past, datasets of this scale were very rare and most likely only available to the research team preparing them. The growth of digital repositories as well as the development of research clouds will enable the Prosecution Project to share its data, initially at least as a dynamic and developing resource. Beyond the conclusion of first stage research funding (under the Australian Research Council Laureate Fellowship scheme) in September 2018 the project will continue under the auspices of the Griffith Criminology Institute, with project researchers continuing to seek funding through other schemes and perhaps in partnership with other international projects. At some future stage the data will be defined as complete as they

can be, archived as such institutional repositories (eg Griffith University, Australian Data Archive), and so become a legacy for those wishing to undertake further research, no doubt with questions and tools not yet envisaged by the current research team. Apart from these wide-ranging benefits to academic researchers, citizen history may also have other ancillary results in terms of encouraging historical thinking among participants (Frankle 2011). Such a possibility has potential significance for society at large in an era that is increasingly critical of a prevailing trend towards ahistorical and short-term thinking (Guildi and Armitage 2014).

While many of the volunteers drawn to crowdsourcing endeavours may already have some level of interest or knowledge about the discipline to which they are contributing, the value of citizen history projects as a form of experiential, hands-on learning in itself has been affirmed by some of the feedback from Prosecution Project volunteers. One particularly active volunteer, for instance, became involved in the PP because of her interest in family history, including that of a convict ancestor. Yet transcribing a large amount of criminal records not only provided her with insights into the justice processes that her early relation would have faced, but a new awareness of the changing sentencing practices that criminal offenders like him were subject to across time. This in turn led to her becoming part of knowledge creation process by co-authoring one of the research briefs on the site (prosecutionproject.griffith.edu.au/crime-across-time-mapping-longitudinal-changes-in-criminal-justice/). Experience in university classrooms too has found that involving students in digital projects enriches their experiences and understandings of history (Alker 2015).

Contributing to projects aimed at investigating and solving social problems can be a personally empowering experience, as Christopher Williams points out in an article reviewing the different ways in which crowdsourcing is now being applied to contemporary criminal investigation and law enforcement (2013). Participation in heritage projects may likewise help citizens confront and better understand troubling and violent aspects of the past,

as well as assisting scholars to better comprehend public perspectives of such subjects (Seitsonen 2017). Equally it has been suggested that crowdsourcing helps build a sense of community by 'democratizing history' (Grove 2011, 6). Ironically, 'crowdsourcing' of historical data may also act as a corrective to the so-called 'extraordinary popular delusions and madness of crowds' by challenging commonly received wisdom about the past, such as that prosecutions for child sexual abuse were rare before the contemporary era (Smaal et al. 2016).

Resourcing and infrastructure for web-based projects

We conclude this discussion by focusing on the resourcing and infrastructure requirements involved in running a web-based research project retieving large amounts of social data, using both researchers and volunteers in the process. It should go without saying that a project designed in this way cannot be undertaken without significant investment of resources, human and capital, to plan, manage and secure the data collection process, including attention to the research outcomes that may flow immediately or in the longer term future. An ongoing issue that has been identified in the proper resourcing of digital humanities projects is their tendency to be funded by one-time investments or grants rather than as part of long-term strategies incorporated into the operating costs of institutions (Grove 2012).

While the Prosecution Project has been enabled by significant external research investment as well as university infrastructure funding, it should be noted that alternative approaches are increasingly available to researchers in the humanities and social sciences. A leading resource providing infrastructure support to those wishing to harness the growing community of citizen scientists is Zooniverse (zooniverse.org/), now nearly a decade old. Its

earlier and largest projects were in the natural sciences, including astronomy and environmental sciences. In recent years a number of social science research teams have deployed the Zooniverse facility to develop their own crowdsourcing projects and recruit volunteers from Zooniverse's existing users. Scripto (scripto.org/) similarly provides an open source tool to enable community transcriptions of document and multimedia files, with plugins available for popular web platforms like Omeka and Wordpress. Projects utilising Scripto have ranged from the transcription of Civil War era letters held by the Newberry Library (publications.newberry.org/civilwarletters) to community involvement in a PhD candidate's attempt to sift through massive amounts of documentary evidence for mentions of the massive complex of underground tunnels constructed by the Nazi officials in 1944 (nazitunnels.org).

Custodians of large collections of data, especially museums and libraries (Cairns 2013; Ridge 2013), increasingly look to their natural community of users to assist them in the process of digitisation, including the tagging of digital products or the transcription of digital archives. Other research enterprises are even more adventurous in their vision of what citizen scientists can bring to big data. The Atlas of Living Australia (volunteer.ala.org.au) is such an example, shared with comparable undertakings in other countries. In addition to acting as a research cloud, encouraging researchers to pool their data sources for the greater benefit of the research community in a wide number of disciplines, the ALA also encourages interested users from the non-research community to participate in data collection by photographing and documenting their observations of the natural world. This site has also now added its own digital volunteer transcription facility to encourage research groups to involve volunteers in the transcription and identification of digital images of earlier scientific collections.

Where to from here? In an imaginative 'prospective retrospective' penned in 2013, Lynn Nyhart depicted an academic culture in 2038 that had advanced significantly as a growing number of citizen science projects – and the efforts of researchers to engage the public with these through experiential learning experiences – led to the development of new research questions and methodologies (Nyhart 2013). Already, the research benefits of crowd-sourcing are increasingly being harnessed to a wide variety of research undertakings. Open source and shared infrastructure solutions are now becoming available to those research teams unable to call on significant local resources within the budgets provided for specific research projects. While the research project described in this article lends itself particularly to the potential of large numbers of supporters to access data from material artefacts of the past, there seems little reason to doubt that in the social world more generally creative design of research projects might in the future add greatly to the collation of research-usable datasets previously unimagined. As the engagement of large numbers of citizens in the cause of scientific research has shown, the social research world will benefit from the new style of inclusion made possible by online technologies.

References

Alker Z (2015) The digital classroom: new social media and teaching Victorian crime. Law, Crime and History 5(1): 77-92.

Cairns S (2013) Mutualizing museum knowledge: folksonomies and the changing shape of expertise. Curator: The Museum Journal 56(1): 107-119.

Causer T, Wallace V (2012) Building a volunteer community: results and findings from transcribe Bentham. Digital Humanities Quarterly 6(2).

Ellwood E R, Dunckel B A, Flemons P, Guralnick R, Nelson G, Newman G, Newman S, Paul D, Riccardi G, Rios N, Seltmann K C, Mast A R (2015) Accelerating the digitization of biodiversity research specimens through online public participation. BioScience 65(4): 383-396.

Finnane M and Garton S (1992) The Work of Policing: Social Relations and the Criminal Justice System in Queensland 1880-1914 Part 1. Labour History 62: 52–70; Part 2. Labour History 63: 43–64.

Finnane M, Piper A (2016) The prosecution project: understanding the changing criminal trial through digital tools. Law and History Review 34(4): 873-891.

Finnane M, Kaladelfos A (2016) Race and justice in an Australian court: prosecuting homicide in Western Australia, 1830-1954. Australian Historical Studies 47(3): 339-363.

Frankle E (2011) More crowdsourced scholarship: citizen history. Center for the Future of Museums Blog: http://futureofmuseums.blogspot.com.au/2011/07/more-crowdsourced-scholarship-citizen.html

Godfrey B, Lawrence P, Williams C A. (2008) The history of criminal statistics. History and Crime. London: SAGE. 25-49.

Grove T (2010) History bytes: the knowledge of crowds. History News 65(3): 5-6.

Grove T (2011) History bytes: citizen history projects. History News 66(4): 5-6.

Grove T (2012) History bytes: citizen history projects. History News 67(1): 5-6.

Guildi J, Armitage D (2014) The History Manifesto. Cambridge: Cambridge University Press.

Hathornwaite C (2009) Crowds and communities: light and heavyweight models of peer production. Proceedings of the 42nd Hawaiian Conference on System Sciences. Waikola, Hawaii, IEEE Computer Society: 1-10.

Holley R (2009) Many hands make light work: public collaborative OCR text correction in Australian historic newspapers. National Library of Australia, Canberra.

Holmes K (2003) Volunteers in the heritage sector: a neglected audience? International Journal of Heritage Studies 9(4): 341-355.

Howe J (2006) The rise of crowdsourcing. Wired, 1 June 2006: https://www.wired.com/2006/06/crowds/

Nyhart L K (2013) The shape of the history of science profession, 2038: a prospective retrospective. Isis 104(1): 131-139.

Piper A, Finnane M (2017a) Defending the accused: the impact of legal representation on criminal trial outcomes in Victoria, Australia 1861–1961. Journal of Legal History 38(1): 27-53.

Piper A, Finnane M (2017b) Access to legal representation by criminal defendants in Victoria, 1861-1961. UNSW Law Journal 40(2): 638-663.

Piper A (2018) Victimization narratives and courtroom sexual politics: Prosecuting male burglars and female pickpockets in Melbourne, 1860-1921. Journal of Social History 51(4): 1-24.

Ridge M (2013) From tagging to theorizing: deepening engagement with cultural heritage through crowdsourcing. Curator: The Museum Journal 56(4): 435-450.

Ridge M (2014) Crowdsourcing Our Cultural Heritage. Farnham, Surrey: Ashgate.

Seitsonen O (2017) Crowdsourcing cultural heritage: public participation and conflict legacy in Finland 4(2): 115-130.

Showstack R (2012) Project to transcribe old ship logs provides important weather data. Eos 93(45): 454-455.

Smaal A, Kaladelfos A, Finnane M. (2016) The sexual abuse of children: recognition and redress. Clayton, Victoria: Monash University Publishing.

Smith M (2014) Citizen science in archaeology. American Antiquity 79(4): 749-762.

Watson D, Floridi L (2016) Crowdsourced science: sociotechnical epistemology in the e-research paradigm. Synthese: 1-24.

Williams, C (2013) Crowdsourcing research: a methodology for investigating state crime. State Crime Journal 2(1): 30-51.