Enabling better data discovery of records across archives, institutions and libraries

Mark Finnane[1], Sarah Nisbet[2], Michael McGuinness[1], Ingrid Mason[3], Nicolas Rossow[1], Malcolm Wolski[1]
[1] Griffith University
[2] eResearch South Australia (eRSA)
[3] AARNet

Making use of a purpose-built structured database with the flexibility to capture data from a variety of different sources and jurisdictions, and enabling a productive partnership between researchers and volunteer transcribers, the Prosecution Project is a platform for revolutionising our understanding of criminal justice histories. The functionality of this database will be extended by operationalising a national, sustainable and scalable API standard that will allow data (and metadata) sharing and transfer between archives, institutions and research projects (such as Tasmanian Archives and Heritage Office, The Prosecution Project, Trove, Queensland State Archives, HuNI and the ALA).

This paper will use The Prosecution Project as a case study to explain how Australian archives are opening up access to resources and show how it is possible for the research community to add value back to the archived material by linking their contributions to the original source material.

Introduction

Historical research is data rich and labour-intensive. Traditionally the data collection methods are paper-based, i.e. documents stored in our systems or print sources available in libraries. Equally, the conventional outputs have been paper-based, largely in the form of published monographs or journal and newspaper articles. Even after the development of computers as a part of the research landscape in the 1960s and 1970s, historians remained tied to paper. Those interested in qualitative analysis typically collected their data on site, in archives offices or sometimes at the original location of the records being accessed. After collection of the data on paper forms, the data might later be entered on a network server through a terminal/application and analysed through dedicated software programs. Connections between the sources of the data and the analytical or written outputs were severed at every stage of this process.

The revolution in personal computing, and the associated development of the World Wide Web in the 1990s, transformed the possibilities of historical research. Yet the take-up of the opportunities enabled by this revolution has been slow. In part, this is a question of the success of tried and trusted methodologies used by historians that have, of course, produced the enormous corpus of historical materials that stock our libraries and constitute a significant record of the political and cultural life of individuals, communities, nations and the international border. The old analogue methods have served the community very well.

But a third element of the digital revolution has been the incremental, sometimes more rapid, development of the possibilities of digitised data. By this we refer to two phenomena. One is the development of digitised photography and other forms of scanning devices that enable the preservation and reproduction of the scanned object as a digital file. The second is the access to the information contained in that digital file in ways that make it possible to analyse the data for a range of search and enquiry purposes.

In contrast to the ready availability of born digital materials that excite the contemporary open data enthusiast, historical materials are necessarily more intractable. Not only may the state of original materials be such as to limit the quality of information available through scanning or photographing a data source, but the current state of digital technologies in any case limits the intelligibility of the data extracted.

This in brief is the landscape that presents itself to researchers interested in taking up the possibilities of the digital revolution to assist their work. Why should they bother? In the case of the Prosecution Project, which is the exemplar for this paper, the reasons are twofold. First, our current understanding of the history of the criminal trial in places like Australia is largely captive to a process of mediation through judicial reports and official statistics, as well as other government sources. Researchers working in this area have long been interested in what can be learned from working much more closely to the original sources, taking samples of materials from the court archives or the police stations and seeing how the process has worked in practice.

Second, historians and others working in the humanities and social sciences have a stake in public education about social life, that may be affected by the emergence of digital technologies and the uses of historical data enabled by them. So, for example, in an area of particular interest to those of us studying the courts and criminal justice in the past, the fact that the sources of information (e.g. court records, police station records) are also the source of information about the lives and fates of individuals, has made these records of very significant interest to a wide range of communities and non-academic users. Evidence for this may be seen since the 1970s in the changing function of public archives, where today one is more likely to find a family history researcher than an academic. Further evidence is also found in the success of commercialised genealogical research (Ancestry, Find My Past, Family Search and so on) enterprises that draw on the records that are used by researchers to trace the history of the criminal trial.

This then is the larger context for understanding the type of research and the possibilities now offered by a digital approach to historical data. This paper will discuss the Australian situation and the current work being undertaken to open up archive resources.

Background (the Australian problem)

To appreciate the scope of this project, it will be valuable to indicate the range of organisations and associations, as well as the research community, that has a stake in this part of the digital revolution.

Organisations

The primary holders of public records in Australia are the public archives or records offices in each state and the Commonwealth. Typically, archives holdings are managed under permissions and other conditions granted by the agencies responsible for the production of the records. Public archives are accessed by a wide range of users, including public servants working for relevant government departments, general community users including family and community historians, and academic and independent researchers. All of these communities value highly the possibilities of access enabled by high-quality scanning and curation of existing holdings. The extent of digitisation of the public records held by the various authorities is not known, but it is difficult to believe that it is at this point more than five to ten percent. Particularly with respect to the records of interest to the Prosecution Project, some archives have been contracting with commercialised search organisations to make available their records to public discovery, but are generally only indexed to the name and date of an entry.

Associations

As indicated above, significant support for the digitisation of public records and other historical materials can be found in a very wide range of community-based organisations. Chief among these are family and local history societies, many of which also supply a significant resource of volunteer labour, that for many years has enabled public archives to extend the quality of their search aids, such as indexes to series records. This resource has been an important part of the Prosecution Project's work, with public volunteers (recruited through the project website and by direct approaches to community organisations) assisting directly in the transcription of the scanned images of court records that form the basis of the project's research.

Another group of users with a significant stake in the expanding digital access world is the academic workforce. It has to be said at this stage the degree of commitment from professional academics to the digitisation and digital research process is something less than its potential. Professional associations of academics engaged in digital research have been active for the last 20 years or so, but the professional associations to which most academics, such as historians, belong have still only a limited engagement in the enterprise. There are important structural features shaping academic work that may be regarded as an impediment to the more rapid development of digital research. They include the prestige associated with conventional publishing, and (conversely) a view that digital technologies and methodologies are no more than instrumental devices with limited research significance.

The records accessed by historians in public archives and manuscript repositories of research and deposit libraries vary greatly in quality of preservation as well as ease of legibility. For modern historians, working anywhere from the eighteenth century, an increasing volume of materials is generally available in printed format. For the kind of records close to everyday life and governmental processes, the great bulk of materials before the twentieth century is likely to be in manuscript form. The capacity of current OCR technologies to render either manuscript or historical printed materials intelligible to the point of returning accessible data through machine extraction is limited. The search capacities of major databases, such as the National Library of Australia's (NLA) Trove digitised newspapers collection, are of course remarkably

good – but still very imperfect. For the foreseeable future, we may speculate that a significant degree of human inspection, rather than technological extraction, is likely to play a major role in accessing the data of the past, for community or research use. One example might be used to demonstrate the point. From the late 17th to the early 20th century, the records of proceedings at the old Bailey Court in London were published in a semi-official format. They have long been regarded as a major source of information about the people who were brought before the court, as well is the processes of the criminal trial and outcomes. It was only with the transcription (by manual double entry) of the original printed proceedings, however, that a research friendly text could be produced capable of being text mined by researchers interested in exploring the larger patterns disclosed by this major record of judicial activity.

Case Study

In Australia a national initiative is underway to improve access to archive data for research and the capabilities of researchers to do so. This initiative is discussed in the following sections.

The Prosecution Project

The Prosecution Project is a major undertaking, based at Griffith University and funded by the Australian Research Council, which has been investigating the history of the criminal trial in Australia. This project has been funded for five years since 2013, with significant infrastructure support from Griffith University, which has enabled the database development being discussed here.

The project aims to conduct research into the history of criminal prosecution across nearly 150 years in all the main jurisdictions of Australia. It does so through the extraction of archival data into a relational database that enables further curation of the records and their preservation for the use of future researchers and other community users. From the beginning, the project proposed to proceed not by way of gathering samples of records, characteristically the pragmatic choice of researchers constrained by time and resources, but through the coordinated collection of data on a comprehensive basis from the relevant institutional repositories. In the project's ambition to collect as much data as can be accessed lies the utility of the database for both researchers and community users. Through a public website, community users can gain access to information collected by the project that is of interest to family and community historians, as well as to those interested in the history of the courts and criminal justice. Through long-term retention of the data in a national research repository, the aim is to facilitate future research in a way not possible before this comprehensive research collection exercise.

The archive materials accessed by the project are typically bureaucratic registers maintained by the court as a record of proceedings. As is well known, the British settlement of Australia and New Zealand made these jurisdictions into somewhat exemplary locations of bureaucratic administration, ensuring the preservation of some very rich and durable record series. An important precedent for the work being conducted by the Prosecution Project, for example, was the major undertaking involving the transcription of Tasmanian convict records into the databases developed for the Founders and Survivors website (Finnane, 2016). For that project, as for the Old Bailey Online, data extraction relied principally on having researchers or volunteers, working in archives repositories, transcribe from the original sources or their microfilm images.

The sources accessed by the Prosecution Project, however, take advantage of the availability of digital images stored on personal or network servers/storage. Moreover, the capacity to stream images to the desktop computers of any transcriber with access to the web has transformed the process of data extraction. Importantly the database constructed from these sources is also linked at the record level to the original source of the data in ways that enable data checking and cleaning, as well as responding to the curiosity of the community user of these images. As the project has progressed from being originally a researcher focused exercise in data gathering in a shared research environment to a new vision of a more open data site of interest to a very large number of users, the potential to extend the utility of the database has become evident. In this context the advantages to the project of an increasing interchange of information between researchers and the institutions that are still our principal data source have also increased.

In summary the question that presents itself as a background to the project discussed here is whether it is possible to construct an interface between the researcher-driven database and those institutional repositories of information about our cultures and communities that are found especially in our major public archives and libraries. This opportunity to develop the interface between the researcher collections and the institutional repository is now the focus of a national project which will use the Prosecution Project as a vehicle to develop relevant standards, workflows and methods to make this happen.

The eRSA Culture and Community Program

In Australia, the National Collaborative Research Infrastructure Strategy (NCRIS) drives research excellence and collaboration between 35,000 researchers, government and industry to deliver practical outcomes. One national project funded under NCRIS is the Research Data Services (RDS) project which has two fundamental objectives related to data services:

1. Data Service development for prioritised research domains/disciplines
2. Continued operational support for existing infrastructure

The Cultures and Communities (C&C) program of work is funded through RDS. It was established in recognition of the need to make both old and new data discoverable, reusable and to extract greater value from existing collections that are as varied as statistical data, manuscripts, documents, artefacts and audio-visual recordings. This national program of activity is led by eResearch South Australia (eRSA). eRSA is a service provider offering expert computing technology knowledge, services and facilities to the research, government and business sectors primarily in South Australia. eRSA has been managing the C&C program of work since 2013 to broaden and deepen researcher access to cultural and humanities data/resources. Typically work undertaken in this domain supports researchers with:

- A data storage facility and registry
- Data management and ingest tools
- Training, workshops and support

The current stage of the C&C program builds upon the successful work of the previous three years' work. The objectives for this stage are:

- Operationalise a national sustainable scalable open API standard that will allow data/metadata sharing and transfer between archives, institutions and research projects
- Facilitate active community participation through the provision of training and education resources for the C&C research community including undergraduate course material, and;
- Provide a consolidated and sustainable user support service.

Historically, archival records have been hidden away in different archives across the country. To access them you need to physically visit the archives, copy the original records and then transcribe the data. With the advances in digitisation and digital archiving solutions, data can be made accessible to the national and international community, and related records can be linked across multiple institutional repositories. The Prosecution Project was a natural candidate to develop a practical working solution to demonstrate an open access approach. This was because of the types of data it was seeking from archives and also the strong working relationship the Project team had built up with the various archive institutions.

The open API initiative is investigating a standardised method of accessing archival data. This goal is to use an API to access to the Prosecution Project's transcription of court records alongside the Tasmanian Archives and Heritage Offices' digitised records. This method of providing data as service, once created, will be able to be exploited by discovery services such as the National Library of Australia (NLA), national and state archives.

To facilitate uptake of the API, the project will also deliver user support and training to enable State Archives, researchers and cultural institutions to easily adopt the Open API standards. This includes:

- Training
    - Engagement team working at a national scale to raise awareness of existing tools;
- User support and documentation work:
    - Development of Humanities User Support Service utilising the Nectar/RDS distributed help desk infrastructure.

An established technical reference group and governance group is determining the best approach for obtaining feedback, validating solutions and promoting uptake within the community. A resource kit for users who want to use the Open Standards will be developed and made available to the community.

The final solution will utilise the National User Support service (https://support.ehelp.edu.au/support/home) established to support Virtual Laboratory's developed through NCRIS funds. This online service will be used to consolidate and develop Tier 0 user support material for the C&C community.

Governance

While eRSA manages the C&C program of work, a Reference Group provides governance. The members of this group are representatives from the various stakeholder organisations from the Humanities and Social Sciences sector to enable broader participation in the Cultures and Communities (C&C) project. The Reference Group also takes advice from a Technical Advisory Group to ensure the project meets the expectations of the C&C community and the various infrastructure/service providers.

Proposed Solution

To date in 2017, the project team has been engaging with a number of key stakeholders, including the National Library of Australia, National Archives of Australia, Tasmanian and Queensland State Archive, as well as the research community. A workshop was held on 24th February which laid out the landscape and potential solutions were discussed.

Figure 1 shows the proposed solution with the state archive at the Tasmanian Archive and Heritage Office (TAHO). The outcome sought is to replace the current manual methods with automatic workflows that result in permanent links between objects held at the two repository/databases.

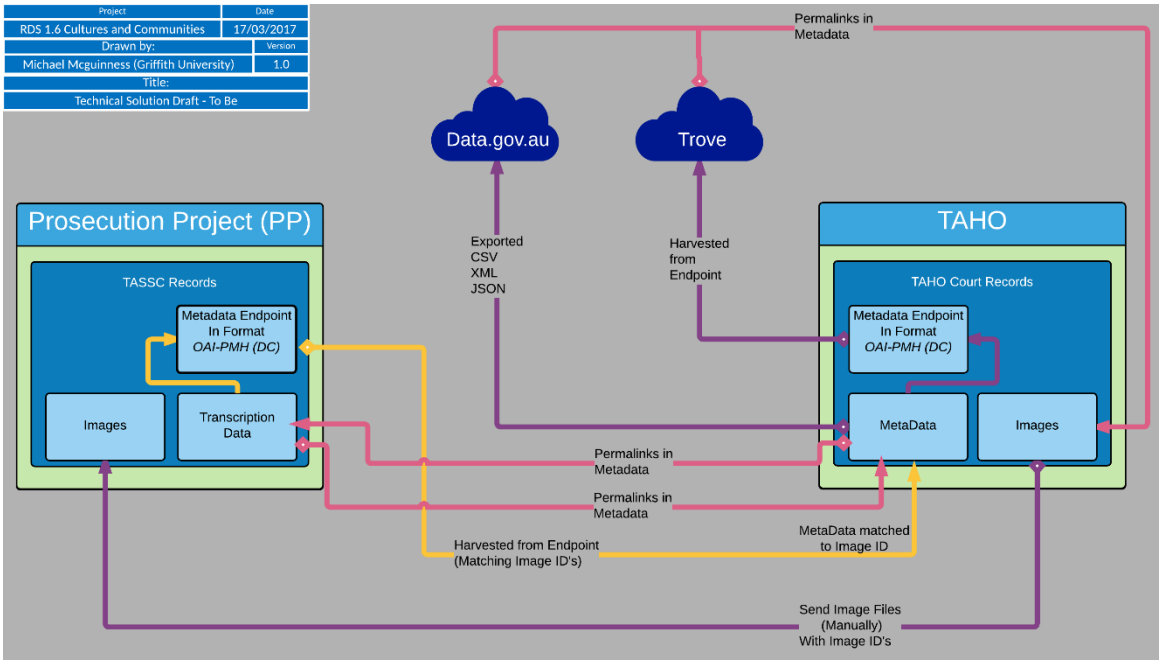

Figure 1 - Open API Roadmap version 1

It is recognised that ultimately there will be a need for the adoption of several methods and standards because of a number of factors, such as legacy infrastructure, leveraging existing skills sets at each endpoint e.g. different archive institutions and research centres etc. The pilot project with the Prosecution Project and TAHO above will develop the first method to test the workflow approach and to define further requirements and specifications for future development activities. The key outcome sought from this pilot activity is to demonstrate that researchers can access archived material and link their work back to the originating source in the archive. This not only makes the researcher's work more accessible to a broader community, but also enriches the original source material in the archives.

There is also future work, out of scope of the current project that has already been identified, that would assist the digital humanities spaces. In particular, the following areas need attention to be able to increase the value of the API to researchers and institutions. These include:

- A Name Entity Recognition workflow (NER) solution
- A connection to a Unique Names Index (UNI) solution
- A Persistent Identifiers (DOI/URI) linked data
- An Acknowledgements (Citations) solution for contributions

Further infrastructure needed

By the end of this development and engagement activity (the RDS project), there will be an open API that is utilisable and has archival metadata that is accessible and reusable by the project partners and a mix of cultural institutions and research groups. A critical examination of a sample of current project inputs reveals considerable terrain for further development and investment. For example:

- There is one data source type (written archival manuscript), yet there are diverse textual documentary history resource types that can support historical research, e.g. printed text, in early newspapers, pamphlets, and periodicals. Image and sound based documentary history resources also operate as data inputs in historical research. The addition of more diverse data inputs would support a comparison of what was recorded as official records in comparison with public and personal viewpoints.
- The metadata schema for the API supports data reuse across diverse cultural institutional types (galleries, libraries, archives and museums), yet semantic requirements are diverse, depending on whether the data reuser works with common or local schematic or entity type data documentation formats. The provision of tools to support user directed metadata crosswalking, with the capacity to publish a range of metadata publishing structures and formats, enables and extends the range and ease of data reuse.
- The technical construction of the API follows common standards, yet there are diverse use cases of data query, e.g. small and highly controlled queries of the dataset or requests for the entire dataset. The provision of different data publishing interfaces, e.g. OAI-PMH, would enable light and heavy weight data requests and reuse.
- The digitised archival content supplied for ingest into the Prosecution Project has a defined scope, yet this digitised material has not been packaged up, such that it can be discovered, reused, and cited as a scholarly data input by other researchers, and found through discovery services such as Trove or Research Data Australia. Coordinated data curation and description of these digitised archives removes the barrier for other historical research to be conducted, and the data is more accessible and reusable.

Critical success factors for the current activity

The critical success factors for the RDS project fall into three areas:

- Stakeholder engagement
- Data access and interoperation
- Technical interoperation

The broader goal of stakeholder engagement has been to foster information and knowledge exchange and build trust to support community building and collaboration, to provide a strong foundation for further research infrastructure development. Project participants share an interest in the project outcomes that serve both humanities research and cultural heritage interests. The project has been structured to ensure stakeholders share their knowledge widely, so that data can be re-used and enriched, and technical resources are widely utilisable. Participant involvement is highly flexible, permitting commitment for technical involvement and uptake to be based on a capacity to provide inputs and receive outputs. The exchange of knowledge around the means for data sharing, exchange, and semantics supports resource and research discovery.

Success criteria for stakeholder engagement are:

- Stakeholders are better aware of the range of requirements that arise through this kind of collaboration.
- Mutual needs for data and technical interoperability can be satisfied through the utilisation and development of national research infrastructure.
- Stakeholders are keen to build on the success of the project, by proposing to work together on a second phase and tackling the data exchange with different digitised cultural datasets and extending the stakeholder base to other cultural institutional types.

The broader goal of data access and interoperation has been to ensure that cultural material targeted for digitisation based on humanities research needs can be delivered back to the institution. Data exchange and

enrichment, are seen as mutually beneficial, thereby creating a virtuous circle. This approach has been taken to address problems of data provenance, data fragmentation, duplication of effort (in digitisation), lack of data reuse, and, semantic interoperability. Lightweight semantic threads will be developed to publish cultural data (maintained in a research platform within the research sector) and link that enriched data back to the source (maintained in collection platforms in the cultural sector).

Success criteria for data access and interoperability will be:

- Flexible options are created that enable wide uptake of the approach to data publishing and semantic structuring in support of diverse needs for descriptive metadata and digitised cultural data.
- Stakeholders who opt to supply and receive data are able to integrate the data into their respective systems such that resource discovery is enhanced and research is enabled.
- Stakeholders within the project or more broadly observing the project outcomes opt to take a similar collaborative approach to enable more cultural data to be made reusable and accessible in support of humanities research.
- Better understanding of digitisation workflows and data exchange practices, spanning the cultural and research sectors, to aid with scoping further research infrastructure and capacity development.

The broader goal of technical interoperation has been to ensure that data contributors reach a consensus on the information standards, data transfer protocols and exchange workflows that are to be embedded in the technical solution. Identifying where technical interoperability requirements are held in common is key to effectively bridging the divergent technical infrastructures maintained in the research and cultural sectors.

Success criteria for technical interoperability will be:

- The technical infrastructure accommodates the technical interoperation needs of participants that provide ongoing institutional resource discovery services or humanities research projects.
- The data contributors from the cultural and research sectors utilise the technical infrastructure developed for data exchange.
- The technical infrastructure developed through the project is reused for further data exchange by project participants or other parties in the research or cultural sectors.

Successful outcomes across all three areas (stakeholder engagement, data access and interoperability and technical interoperability) in this project will ideally inform the nature and scope of any shared national research infrastructure developed at a greater scale.

Conclusion

Drawn together, the above work and assessments of needs to date lead to a bigger question about the greater challenge in national research infrastructure for the humanities:

How can this project scale into a national platform, such that many more data driven humanities research projects and cultural institutional partners can repeat and benefit from this type of collaboration, and, share infrastructure?

The process undertaken in this RDS project is typical of all digitisation arrangements made between the keepers and users of cultural assets in analogue format. When material in cultural institutions is identified by humanities researchers as a critical input for their research, key steps follow on from this: (1) Liaison is undertaken and planning occurs to jointly scope the digitisation activity; (2) Funds are sought for digitisation and support services; (3) The transfer of material in the digitisation between: the service provider, the institution and the research recipient, is arranged. There is an obvious pathway for digitised material to be uploaded and to flow between these parties that is yet to be effectively exploited as a data supply chain in service of research. Many groups contribute to the effectiveness of this supply chain. For example, the research and education network provided by AARNet is designed to support secure large scale data transfer and it is provided to the three types of organisations in this supply chain: universities, service providers, and cultural institutions.

If the outputs and outcomes of this project demonstrate sufficiently that a data supply chain and a virtuous

circle can be established through partnership and collaboration, there is an additional question about exploring options to scale this up to being a national data supply chain for all research, and not just the humanities. This would require investment in the support for the collaboration and technical services to be coordinated into the use of the research network, cloud storage and transfer, and platforms that underpin Australian national research infrastructure. Investment in the efficient use of infrastructure resources already available to these collaborating parties, through the development of standard processes for collaborative digitisation arrangements, would expedite the data flow. Further analysis and effort could be applied to link national and institutional infrastructures, such that they operate in tandem and service both high end research, and community needs, for access to large cultural datasets. Australia's advantage with many of the stakeholders operating on the national research and education network, unlike other nations, is yet to be leveraged effectively.

References

Finnane, M, Kaladelfos, A, Piper, A, Smaal, y, Blewer, R and Durnian, L et al The Prosecution Project Database https://prosecutionproject.griffith.edu.au/prosecutions (version 1, 17 July 2016).

Klingenstein, Sara, Tim Hitchcock, and Simon DeDeo. 2014. "The Civilizing Process in London's Old Bailey." Proceeding of the National Academy of Science. http://www.pnas.org/cgi/doi/10.1073/pnas.1405984111.

Finnane, M. 2016. "The Prosecution Project: Investigating the Criminal Trial in Australian History". *Humanities Australia*, Vol. 7. pp. 35-45. http://www.humanities.org.au/Portals/0/documents/Publications/HumanitiesAustralia/Issue_7/HumanitiesAustralia-07-2016-Finnane.pdf

Mark Finnane is ARC Laureate Fellow and Professor of History at Griffith University. He has published widely on the history of criminal justice, policing, punishment, and criminal law. He directs the ARC-funded ‘Prosecution Project’ , hosted at the Griffith Criminology Institute.