DOING RESEARCH DIFFERENTLY

ARCHIVING & SHARING QUALITATIVE DATA IN STUDIES OF CHILDHOOD, EDUCATION & YOUTH

JULIE MCLEOD, KATE O'CONNOR, NICOLE DAVIS

Doing Research Differently: Archiving & Sharing Qualitative Data in Studies of Childhood, Education and Youth

Authors: Julie McLeod, Kate O'Connor and Nicole Davis

Contributors: Sari Braithwaite, Emily Fitzgerald, Rachel Flenley, David Haddican, Jo Higginson, Janet McDougall, Monika Popovski, Henry Reese, Geordie Zhang.

Doing Research Differently: Archiving & Sharing Qualitative Data in Studies of Childhood, Education and Youth

Publisher: The University of Melbourne

DOI: 10.25916/5e9e28eec21a1

www.socey.net

Cover image: Petrie Terrace State School, Brisbane, May 1970. Queensland State Archives, Digital Image ID 25777

TABLE OF CONTENTS

1. INTRODUCTION

> *I am not arguing that my third reading is 'truer' than my first or second; they are, simply different.*[1]

Internationally there is growing enthusiasm and expectations for researchers to make their research data openly available for use by other scholars and interested parties, including the wider public. Funding organisations are increasingly introducing requirements for the publication of datasets and actively encouraging data sharing; concurrently, institutions and researchers are looking at new ways of storing, managing and disseminating their research data. There is a corresponding interest in experimenting methodologically with data sharing and addressing its ethical, methodological and practical challenges. Overall, researchers, institutions and funding bodies are giving greater attention to documenting best practice in the storage, management and accessibility of data for re-use or secondary analysis, taking account of regulatory settings and opportunities for innovations in how research is undertaken and communicated.

Secondary analysis of social science data has been more typically associated with re-analysis of quantitative datasets, either using a new technique of analysis or asking questions on a topic that was not part of the original study's main focus (e.g., Smith 2008). Some protocols for qualitative re-analysis can be learnt from the experience of quantitative secondary analysis, but there are obvious and well-documented differences in the type of data and process for eliciting or collecting data, in negotiating the context of the original study and in how data are recontextualised in any subsequent interpretation (Heaton 2004, Fielding 2004).

This discussion paper explores directions and dilemmas in the archiving and sharing of qualitative research, taking a specific focus on studies of childhood, education and youth, predominantly from across the social sciences. It was prepared as part of a program of work funded by the Australian Research Data Commons (ARDC), which comprised the development of a pilot qualitative data archive and sharing platform: SOCEY (Studies of Childhood, Education and Youth), and the archiving of data from six projects within that archive. The paper is intended as a supplementary publication to more extensive general guides such as *Managing and Sharing Research Data: A Guide to Good Practice* (Corti et al. 2014) and focuses on the Australian context. While there is a growing international body of work on re-using qualitative data (Grant 2015), to date there has been comparatively modest engagement with these matters among the social science research community in Australia.

We begin by discussing the opportunities and challenges for archiving and data sharing in qualitative research (Section 2) and provide an overview of Australian and international examples of archiving and sharing (Section 3). We then detail the development of the SOCEY pilot (Section 4), before considering protocols and exemplars of best practice for archiving and sharing research data alongside the experiences of those who conducted the pilot archiving (Section 5). Finally, we propose some key principles intended to inform our future work in this area and define our next steps (Sections 6 and 7).

[1] Riessman 2004, 321. Quoted in Andrews 2011, 7.

2. ARCHIVING & SHARING QUALITATIVE RESEARCH DATA: OPPORTUNITIES AND CHALLENGES

2.1. What is Digital Archiving?

Digital archiving in the context of academic research is the storage of research data in a digital environment (online or offline) – a research data repository (University of Sheffield n.d.). In qualitative research, this involves a wide range of media, both born-digital or subsequently digitised. Some of these media include video or audio interviews, either born-digital or digitised from analogue formats; born-digital transcripts of such interviews; scanned participant consent forms; spreadsheets of permissions and participant details; photographs; fieldwork notes and more. Research data might be included in a repository with no access beyond the immediate researchers or may be intended for sharing with present or future use by others. Data repositories might also be used by researchers to store data in a way that would leave open the possibility for access to the data to be made available after a certain period of time has passed. In whatever ways in which they are developed and used, 'research data is best preserved and published using a research data repository' (University of Sheffield n.d.).

2.2. What is Data Sharing?

Data sharing can involve different forms of data and access conditions. It can include published and processed datasets that underpin a publication and have their own DOI. Data sharing might mean completely open access to 'raw' unprocessed data – accessible to all with no restrictions from the time it is loaded to a repository – or it might mean that access to data is mediated by the lead researcher or repository staff with specific access restrictions in place. These restrictions may include open access to some materials but not others; timeframes for restrictions on specific materials or entire collections; and access only to anonymised data, to name a few. There are decisions to be made by researchers as to the mix and format of raw or unprocessed data they might archive, notwithstanding the complexity in making distinctions about this in relation to qualitative research.

2.3. Drivers for archiving and sharing

There are multiple drivers for archiving and sharing qualitative data, including from funding bodies, governments, other scholars and expectations from the public for greater accessibility and transparency of research. In Australia, one of the largest funding bodies, the Australian Research Council (ARC), has built data management and sharing requirements into their codes and guidelines, as well as funding applications. These include making research outputs openly accessible, creating a data management plan, and encouraging 'appropriate access by the research community' (ARC 2019).[2] Beyond matters of compliance, a range of other parties have an interest in data sharing, such as researchers in the similar or complementary fields, community

[2] Research outputs have, since 2013, been required to be 'made openly accessible within a twelve (12) month period from the Publication Date', with certain exceptions and restrictions (ARC 2017). The ARC is also committed to making research data more accessible, 'encourag[ing] researchers to deposit data arising from research projects in publicly accessible repositories' (ARC 2019). Data management plans must also (since 2014) be built into applications, in line with the Australian Code for the Responsible Conduct of Research 2018 published by the National Health and Medical Research Council (NHMRC), which created these guidelines with qualitative data in mind but which can also be applied to qualitative data (ARC 2019; NHMRC 2018).

and advocacy organisations, NGOs, and government departments. The Australian Data Archive (ADA n.d.) lists two principle reasons for depositing data in a repository: to ensure preservation of research data for the future and 'to enable it to be shared with others for secondary analysis'. At a 2019 Roundtable hosted by the Studies of Childhood, Education and Youth network (See Section 4.2 below), representatives from local, state, and federal government departments, NGOs, and other community and lobbying organisations focussed on children and youth saw great value in having access to completed and processed qualitative research data that offered them access to the perspectives and voices of participants. They saw this as allowing them, for example to, draw on lived experience data and narratives to inform approaches to policy and the development of programs and practices.

2.4. Opportunities

An important question for qualitative researchers, and a particularly pressing one for studies of childhood, youth and education, is how to build upon and aggregate insights from previous studies; and, then, how to maximise the value of the rich and in-depth sources that are created in the course of doing a qualitative study. Typically, the sources and material generated in qualitative studies – field notes, transcripts, case profiles, coding schema, preliminary analyses, question schedules, and visual and other artefacts – are regarded as being solely for the private use and interpretation of a researcher or research team. There are strong reasons for this, discussed in Section 2.5 below. However, the convergence of developments in digital technology, increasing expectations for data sharing deriving from funding bodies and government agencies, and growing interest in the value of archived data for secondary analysis are challenging these views (Corti 2000; Heaton 2004). The digital revolution has enabled the development of novel approaches to and sources for conducting research, and has transformed opportunities for storing, representing, disseminating and re-analysing research on a scale unimaginable for most social science researchers even a decade ago. Such possibilities for data sharing and re-analysis have begun to unsettle established ideas about the relationship between researcher, research context and data, with mixed effects (Bishop 2009; Fielding 2004; Parry & Mauthner 2004; Crossen-White 2015).

Despite these dilemmas, data sharing nevertheless offers many benefits for qualitative sociological research. It opens up possibilities for transparency in the practices, methods, and outcomes of educational research, and has the potential to enhance rigour and impact. Greater access to the records of current qualitative projects provides 'archives for the future' and supports historically-informed approaches and sensibilities in sociological research. Revisiting data from earlier qualitative research projects not only offers comparative perspectives on particular social phenomena, it also affords a historical perspective on the methods of researchers and the history of social science practices. Re-engaging with qualitative data from past research projects is likely to be a more familiar methodological tool for historians and, particularly, oral and social historians, who have long returned to previous studies (along with other historical data) to inform new research (Dening 1998, Crossen-White 2015).

In his study of mobility and social change in postwar Britain (2010), sociologist Mike Savage has argued that re-use of research data offers new insights into the history of expertise and the social science disciplines, including how knowledge about social change is produced. *Making Futures*, one of the studies that forms part of the SOCEY pilot project discussed below, developed a methodology that linked analysis of approaches and data from earlier major studies of youth and schooling to the contemporary study so as to document the history of expertise about young people. Andrews (2011) has reflected on how she revisited her own qualitative research on British social activists several times over a period of twenty years, gaining new perspectives (both from the data and her own life experiences) on child-parent relationships and aging from material originally gathered to examine 'sustained political commitment'. A further example of potential

benefits in re-engaging with archived data is found in the UK project 'Reanimating Data: Experiments with People, Places and Archives', which is re-engaging with research data to bring it to life in new and creative ways that show the life and value of this material beyond formal and institutional research settings (Reanimating Data n.d.). This is a 'collaboration between academics, archivists and activists interested in young women's sexual health and empowerment … working with a set of interviews collected as part of a social research study conducted in Manchester in 1988–90.' The project aims to 'archive, share and reanimate this material as way of exploring change and continuities in intimate lives over a 30-year period'. This work is developing in conjunction with artists and members of the community where the original research was undertaken: outputs include sound installations, linking with former participants and groups and returning datasets, and connecting academic and community archives.

From a future perspective, particularly looking through the gaze of an historian or sociologist, there is significant value in being able to identify the experiences and perspectives of participants and the context in which they lived, in addition to understanding the context of the original study itself. Such work can open up possibilities for doing qualitative research differently, for fresh thinking about its value and purpose in the present, and for anticipating its value in the future as an historical source for understanding social change.

Exposure to the methods underpinning research projects also brings significant educational benefits for teaching and mentoring new and emerging researchers. It can show how project design and aims are realised methodologically and then how these ideas are put into research practice. In doing so, it reveals the iterative process of qualitative analysis and the nature of qualitative methods in action, providing valuable exemplars for novice researchers.

2.5. Challenges

Creating digital archives in order to support data sharing and re-use gives rise to specific ethical and epistemological challenges. This is particularly so for qualitative research, where issues of context, purpose and specificity of data, participant consent, and aims of re-use are significant considerations. For example, managing participant consent and any undertakings of anonymity, ensuring that the context and specificity of the original research is not lost or diminished. Does consent for the earlier project endure over time? How are participants fully informed about intentions to share data at the end of the project? How to ensure that archived material is presented and curated in a way that does not imply data can be neatly abstracted from the context of its production? How do researchers feel when their data is interpreted in ways they might never have anticipated or consider to be mistaken? (Broom et al. 2009; Hammersley 2010; Moore 2007; Andrews 2011; Crossen-White 2015). Many of these matters are amplified in digital environments, with the potential for the rapid and decontexualised online dissemination of data especially when appropriate protocols and protections are not in place; and, even if such protocols are in place, they might not be sufficiently robust or regulated. A further significant consideration in settler-colonial Australia concerns the collection, curation and sovereignty of Indigenous data, and how this is stored, accessed, and 'made open' in museums and other public institutions.[3]

The increasing emphasis on digitisation, sharing, reuse, and mediated or fully open access to research data also complicates traditional notions of how we treat people's data and stories in research projects. Such discussions have been debated among oral historians for some time (e.g. Bornat 2003, Andrews

[3] Increasingly, particularly in the GLAM (galleries, libraries, archives and museums) sector, Indigenous participation with and co-curation of their own archives and stories is being adopted as best practice. See, for example, Daniels & Senior 2019; Callison, Roy & LeCheminant 2016; Thorpe & Galassi 2014. See also the Indigenous Data Network at Melbourne School of Population and Global Health n.d.

2011, Thor Tureby 2013, Bornat 2013) with many of these considerations resonating with the methodological and ethical questions raised by the SOCEY pilot project.

One of the key arguments against the re-use of qualitative data is that the material is culturally constructed, co-produced by the researcher and researcher participants, and cannot be properly understood outside its original purpose, context and conceptual and empirical framing (Hammersley 1997; Mauthner et al. 1998; Broom et al. 2009; Andrews 2011). Yet, as Broom et al. (2009) argue, 'even primary data are contextual, partial and incomplete, thereby creating contextual problems even for the original researcher'. In addition, the original researcher reinterprets their data after it has been created by producing outputs from that data – sometimes multiple times and often many years after that data was originally created. A related concern from researchers goes to the heart of who owns the data and the *ownership* of knowledge built from that data; there can be a sense in which the original researcher feels they 'own' the data, or it belongs to the participants and some reluctance for it to be appropriated by others or analysed in other ways.

In summary, key ethical, methodological and epistemological challenges include:

- managing the re-use of qualitative research materials without compromising the specificity of the context in which they were produced;
- creation of appropriate contextual materials to guide re-analysis of archived qualitative datasets;
- transparency and care in obtaining participant consent for archiving and use of research materials at the time of data collection as well as subsequent re-use;
- ethical, regulatory and practical protocols governing the management of access to archival repositories;
- identifying appropriate ways to mitigate perceived and/or actual risk to participants
- recognition of the impact on the original researcher or research team of subsequent engagement with their data by others
- recognition of historical abuses of data access and appropriate protocols in accessing Indigenous data
- distinguishing between data sharing as a matter of policy and funding body compliance and as an opportunity for creative methodological innovation;
- decisions on determining on what to archive and how to organise materials; and
- accessibility (e.g., provision of appropriate metadata, access to data and outputs) for non-academic stakeholders, including NGOs.

While the importance of these challenges cannot be underestimated, they are not necessarily in-principle reasons against archiving and re-analysis. Rather, they are cautions that warrant further investigation and careful and iterative evaluation of evolving practices. As we discuss in the next section, concerns about the future use of research data could be mitigated by exemplars of responsible practice and by documenting researcher-led processes, such as those undertaken in our pilot study.

2.6. Designing research for data archiving and sharing

Archiving a qualitative study involves far more than simply depositing research data into a physical or digital repository. Decisions need to be made about how much contextual and biographical information is necessary to enable meaningful access and re-analysis, while not compromising ethical agreements. This may involve further interpretive work in deciding how to tell the story of the study, what matters most about the findings, and what the study might contribute as a resource for future research (Andrews 2011). Although permissions and protocols can be renegotiated and re-evaluated during and after the study is complete, ideally a comprehensive plan would be well established prior to beginning a project. In this way, both

researchers and participants understand what will and may happen with the data, both for immediate research purposes and any future research.

Academic institutions and many funding bodies increasingly require research proposals to contain sound data management plans, as well as where relevant human ethics approvals. Previously a standard and acceptable practice for qualitative researchers was to retain their data – recorded interviews, transcripts, and other materials – either in their own possession or perhaps a physical archival space within their institution. Participants' data would have typically been drawn upon in outputs such as academic journal articles, monographs or government reports, usually de-identifying participants and their broader communities in order to preserve their anonymity. In addition, for some studies, both qualitative and quantitative data might only be retained for a limited period of time and then destroyed, protecting the participants but negating any future usefulness of the data.

The expanding digital realm requires additional considerations for data management and potential re-use of data, which to an extent are driven by government and university policies on data sharing. Careful planning and 'future proofing' in data management plans helps ensure that initial researcher/s have appropriate structures and permissions in place to support responsible data archiving and re-use where desired. In addition, such early stage planning allows study participants to better understand and indeed have input to decisions about whether and how their data might be used in other projects.

Many ethical and methodological concerns regarding qualitative research data could be allayed by following guidelines designed for the management of research data. Useful examples include the *Management of Data and Information in Research* developed by the NHMRC in Australia for quantitative data, as well as international model guidelines on the archiving and sharing of qualitative data (e.g., UK Data Archive [UKDA], Qualitative Data Repository at Syracuse University [QDR]). While guidelines alone are not sufficient to ensure ethical and responsible practices, they are an important starting point and basis for broader discussion and education. The Australian and international contexts for the archiving and sharing of qualitative data are discussed in below, before moving to discuss current guidelines.

3. ARCHIVING & SHARING OF RESEARCH DATA: AUSTRALIAN & INTERNATIONAL PERSPECTIVES

3.1. The international context

Research data repositories that have the aim of both archiving and sharing quantitative and qualitative data are growing in number. Some of these are institutional repositories, designed for those at specific universities to deposit and share their research material, others have a focus on particular academic disciplines (such as the UK Archaeology Data Service [ADS], while others still have a broader remit and are open to researchers from any discipline or institution or the wider research community (e.g., UKDA).

As noted above, there has been more rapid and widespread uptake of these resource in relation to quantitative data. Some of these repositories accept mixed methods research data – both quantitative and qualitative (e.g., UKDA) – but, by and large they have been focused on the former. In recent years, however, there have been concentrated efforts to design repositories

specifically for qualitative data, so as to more appropriately attend to the unique nature of this material.

Some of the methodologies and guidelines, including questions of ethics, that apply to the quantitative data field can potentially be adapted to qualitative material. But the latter have unique considerations, as discussed in Section 1. Not least of these is the varied types of data that might be stored or shared; the large amount of anonymisation that may be required; and the need to ensure that the research is not naively abstracted, and remains grounded in its cultural and temporal context for any future engagement or re-analysis.

The need for repositories that cater to these unique aspects of qualitative research data has led to the development of a number of specialist repositories. These include the UK Data Archive (UKDA), the Inter-university Consortium for Political and Social Research (ICPSR) at Michigan University, and the Qualitative Data Repository (QDR) at Syracuse University and the UK Data Service (UKDS). These international repositories provide some excellent models for Australian efforts to establish qualitative data repositories and develop appropriate guidelines and practices.

3.2. The Australian context

In Australia too there is a recognised need for research data repositories in which both quantitative and qualitative material can be archived appropriately. While drivers for data archiving and sharing come from government, research funding councils and institutional policy, as discussed above, there is also great interest among the wider community in archiving and sharing data. This community is not limited to academic or institutional researchers, but also includes those working in organisations and governments that seek research to inform policy and practice (see below on the case study of SOCEY engagement with end-users).

While some Australian universities maintain their own repositories, the largest and most successful nationwide repository is the Australian Data Archive (ADA). Established in 1981, the ADA is a primarily quantitative data archive that acts as a repository for over 6000 datasets from more than 1500 projects and studies dating from 1838 until the present, and largely social sciences focused (ADA n.d.). Data collections range from historical census data to contemporary longitudinal studies, with data deposits from research conducted by Australian and international government, academic, and NGO contributors.

The long-held desire for a dedicated qualitative research repository in Australia is demonstrated by the development of the Australian Qualitative Data Archive (AQuA) 2007 to 2009 (Cheshire, Broom, & Emmison 2009), funded through the Federal Government's National Collaborative Research Infrastructure Scheme, alongside the establishment of other archival services all operating as part of a distributed ADA network. AQuA began with a review of provisions for the archiving of qualitative data and consultations with researchers to ascertain levels of interest. At that time, this endeavour faced some difficulties gaining traction among the broader social science community. This was in part due to the challenges noted above and, in part, to different policy climates and uncertainty of ongoing funding. It is important to acknowledge these reservations and concerns in developing new initiatives and approaches to archiving and re-using qualitative data. Building a community of practice and actively engaging researchers in these processes is essential. It is with these and other considerations in mind that a group of Australian researchers launched a pilot project for a qualitative data repository grounded in the field of childhood, education and youth studies.

4. SOCEY: STUDIES OF CHILDHOOD, EDUCATION AND YOUTH

The SOCEY project is a small pilot project that aims to provide researchers in the interdisciplinary field of childhood, education and youth studies with a platform to both deposit and share their data and to lead methodological and ethical debate on these matters. As part of this, it is developing an Australia-wide research network and community of practice in which ideas, guidelines, responsible practices and innovation in data sharing and archiving can be advanced.

4.1. Background

SOCEY initially began in 2018 with a core team based at the University of Melbourne and the Australian Data Archive (ADA). Its threefold aims were to:

- provide a platform to connect those working in the interdisciplinary field of childhood, youth and education studies, encourage discussion about research and ideas, and showcase the projects, programs and people;
- explore debates and possibilities surrounding qualitative research sharing and archiving in the humanities and social sciences, particularly in interdisciplinary studies of childhood, youth and education; and
- develop a qualitative data repository, tailored specifically for studies of childhood, youth and education.

The first aims of the project were to develop a website that provided such a platform for showcasing projects and discussing qualitative research archiving and sharing, as well as to provide a portal to the data repository being developed concurrently. The website, http://socey.net, launched in mid-2019 and, as of December 2019, features ten projects from across Australia, with profiles of SOCEY community affiliates.

At the same time, the team began developing the data repository, hosted by the ADA alongside its main data repository, which is housed on Dataverse open source data repository software (Dataverse n.d.). Since 2018, the team has been depositing a number of initial projects, which aim to both populate the repository, as well as provide test cases to assist in the development of guidelines and protocols for the deposit of future qualitative research data.

4.2. Collaboration and consultation

During the development phase of the website and repository (2018/2019) we have also been engaging with numerous stakeholders Australia-wide, both within and beyond academic communities.

In August 2018, a two-day seminar at the Melbourne Graduate School of Education, involved academics from across the country who work in the field of childhood, education and youth studies, as well as those interested in questions regarding the archiving and re-use of qualitative data. This led to the initial core group of scholars now part of the SOCEY community and whose projects are featured on the website. Based on case study presentations of projects, the productive discussions at this event considered conceptual and methodological questions, practical and

ethical challenges for qualitative data archiving, sharing and re-use, as well as imagining what a repository for qualitative data might look like and what it might offer researchers and other end-users. A key question was whether and how this archived material might be used by others, no matter how appropriately the data was stored and curated.

A second consultation took place in September 2019, this time a roundtable with invited representatives from potential external end user groups involved in work on children, youth and education to provide feedback on the SOCEY project. This included participants from municipal, state and federal government bodies, as well as NGOs and advocacy groups such as the Brotherhood of St Lawrence, The Smith Family and Save the Children – all of which work with and/or develop policy for children and youth in Australia. The SOCEY Roundtable invited these groups to engage with the project and the repository, including providing valuable feedback on how these could be of benefit to their work, and noted the additional resources and explanatory tools required to maximise use (O'Connor 2019).

Many of the participants expressed an often-felt divide between their own work and academic research on childhood, education and youth, particularly because many organisations faced difficulties in gaining access to large research libraries and current outputs and data from relevant research, which are often held within databases and publisher repositories to which they do not have access. The group overwhelmingly supported the extension of the project, expressing much interest in being able to both access data in the repository, provided there was appropriate metadata and contextual material to allow easy access for researchers often working under tight timelines. They were especially keen to access the distinctive features of qualitative data which can provide insight into lived experience and the voices and perspectives of diverse communities of children and young people. There were also some initial discussions about the potential to link the SOCEY website and ADA repository to their own datasets in order to promote sharing of research data across sectors and organisations, opening up further opportunities for greater engagement and collaboration between university-based researchers and external organisations.

4.3. Current scope

SOCEY is intended to be an Australia-wide multidisciplinary community of practice in the broad field of childhood, education and youth research. It seeks to provide both a platform for showcasing people and projects through its website, a place for discussion of ideas surrounding qualitative research archiving, sharing and re-use, and a repository for qualitative data accessible for no cost to those wishing to deposit or access data.

While initial development came from a University of Melbourne team, the community involves participants and features projects from a number of institutions including Australian Catholic University, Australian National University, Deakin University, La Trobe University, Monash University, and the University of Melbourne. Scholars come from a range of disciplines in the broader social science field, including sociology, education, history, and health.

A number of these projects represent multi-institutional linkages, often funded by the ARC and other bodies. Some also involve groups from outside the academy, including the Multicultural Youth Australia project, which is a collaboration between the University of Melbourne and nine partner organisations, including government bodies and NGOs that work with and develop policy for youth (Multicultural Youth Australia n.d.).

SOCEY also actively invites participation from groups outside the academy that work in this field – i.e., that conduct research projects, produce reports, develop policy and work with children and

young people. Not only could such groups benefit from access to the research available through the SOCEY Repository but would also be able to deposit their own data for archiving and sharing. With these groups only recently involved in the discussion, we hope towards the beginning of 2020 to start featuring and hosting such projects on both the website and repository.

4.4. Funding

Initial funding for the project came from the Australian Research Council through a research grant to Professor Julie McLeod and the *Making Futures* project,[4] with in-kind assistance from the ADA, SCIP, and the Humanities & Social Sciences Data Enhanced Virtual Laboratory (HASSDevl)/Tinker, which hosts the SOCEY website (SOCEY 2019).

Further support has come from the Australian Association for Research in Education (AARE), which allowed us to begin website development and host a two-day seminar on the project (McLeod, O'Connor & Davis 2018).

Most recently, SOCEY has been the recipient of a significant grant from the Australian Research Data Commons (ARDC), which has provided the funds for progressing the website and repository, administrative support, the archiving of six test projects, the development of comprehensive guidelines, and the 2019 SOCEY Roundtable, inviting stakeholders outside the academy to contribute to and participate within the community.

Currently SOCEY funding for administration, further research, maintenance of the website, marketing and communications does not extend beyond 2019, with further funding being sought for 2020. However, the project does have the advantage of committed support from the ADA, which has ongoing funding, and the SCIP platform at the University of Melbourne.

4.5. Test projects

SOCEY was developed specifically to provide a platform for the archiving, sharing and re-use of qualitative research. As such, the pilot project has involved preparing a number of selected projects for archiving. These were both large and small projects, conducted by researchers ranging in experience from high level academics to graduate students. The projects included:

1. *Schooling Memories: Educating the Adolescent, 1930s–1970s*, a collection of oral history interviews with former teachers, students, curriculum personnel, guidance officers and school counsellors who were at school in the decades from the 1930s to the 1970s, conducted as part of a broader Australian Research Council Discovery Project (led by Julie McLeod and Katie Wright, archived by Emily Fitzgerald)
2. *Globally Mobile Lives and Suburban Government Schooling*, a doctoral research project comprising interviews with secondary school students with experience of work and education across different national contexts and their parents (led and archived by Joanne Higginson)
3. *Childhood Maltreatment and Public Inquiries*, an historical sociological ARC DECRA project that examined the unfolding of the Royal Commission into Institutional Responses to Child Sexual Abuse alongside past Australian inquiries, comprising interviews with advocates of abuse survivor groups, commissioners and staff members of Inquiries, and journalists (led by Katie Wright, archived by Sari Braithwaite)

[4] Australian Research Council, *Youth Identity and Educational Change since 1950: Digital Archiving, Re-using Qualitative Data and Histories of the Present* (FT110100646, Future Fellowship, 2012–2018, Julie McLeod). See also *Making Futures* n.d.

4. *The Curriculum Policies Project*, a collection of interviews with 17 state curriculum experts and education policy makers conducted as part of a broader Australian Research Council Discovery Project (led by Lyn Yates, archived by Henry Reese)
5. *Making Futures*, a longitudinal interview-based study of young people in the final years of their schooling and their parents conducted as part of an Australian Research Council Future Fellowship (led by Julie McLeod, archived by Monika Popovski)
6. *Our Lives Asylum Seekers*, comprising summaries of interviews with young people in Queensland conducted as part of a doctoral research project (led by Zlatko Skrbiš and Jacqueline Laughland-Booÿ, archived by Rachel Flenley and Henry Reese)

Each archiver created a document that tracked the all steps they followed to prepare the data, including referencing all the documentation on which they based their decisions (See Appendices 1–6). In addition to the test projects, the team also explored software for automated anonymisation of data (led by Geordie Zhang).

The team followed ADA's guidelines: their online guides, *Depositing Data* (ADA 2018) and the ADA online *Self Deposit Guide* wiki (2019), which cover both qualitative and quantitative datasets; and the 2014 *Archiving Qualitative Data from a Completed Project* (ADA 2014), which draws on and refers back to other repositories guidelines, such as those mentioned above. The ADA has also been developing a step by step deposit guide for qualitative data deposit in the SOCEY Repository, which will be informed by this discussion paper.

In the following section, we discuss models and best practice for data sharing and archiving, drawing on our experiences with the SOCEY pilot.

5. BEST PRACTICE PRINCIPLES FOR DIGITAL DATA SHARING AND ARCHIVING

As Bishop (2009) discusses, issues of consent and confidentiality can seemingly be barriers to researchers when considering archiving, sharing or re-use of data. In the past, concern has centred on a number of questions, including:

- Should research data be archived? This might be of particular concern if the researcher is considering archiving older materials.
- What are the parameters necessary when archiving surrounding availability of confidential data and anonymisation of the research material?
- Who should have access to this data in the future and how is this mediated?
- What materials should and should not be included in the data deposited and available for sharing.

With the development of numerous quantitative data repositories over the past few decades, and the increasing number of qualitative data repositories, these methodological questions have been much debated. Of particular concern within many of these model guidelines and the broader conversation about archiving, sharing and reusing qualitative data are:

1. Issues of consent and confidentiality
2. Decisions surrounding anonymisation of data
3. Setting access conditions and various levels of access to data
4. Selecting, preparing and formatting materials for deposit

This section will address these four main points, taking account of the detailed guidelines provided by current models and informed by the experiences of the SOCEY pilot.

5.1. Consent and confidentiality

Pre-planning how data will be managed, including archiving the project at its completion, and considering the possible future uses of this data (including for websites, research outputs and more) assists both the researcher and the participant to know what research data will be used for and how it will be handled. It can also alleviate the need to revisit permissions in the future, perhaps many years after the initial consent has been given by participants and they may be difficult or impossible to track down.

The Australian National Data Service (ANDS 2018) in *Data Sharing Considerations for Human Research Ethics Committees* provides a sound base document for researchers to consider when designing research projects and obtaining permissions for archiving, sharing and re-use from their participants. This document is based on Australian legislation that might affect the sharing of research data, including the *Privacy Act 1988*, the *Australian Human Rights Commission Act 1986*, and the *Freedom of Information Act 1982* (ANDS 2018).

Some of the main principles of consent in data sharing (and archiving) from the ANDS guide include (italics added):

- Researchers must gain *informed consent* regarding individual's participation in the research itself and the use of the information that is collected.
- Data, including personal and/or sensitive data, must remain confidential, unless the participant has given permission for these to be disclosed.
- Researchers should 'still exercise judgement to preserve the interests of the participants' and they should not share the data 'if a particular risk is identified' that might 'reasonably, have caused a participant to decline consent'
(ANDS 2018, 6).

Thus, when obtaining consent for research the participant should be informed about and the possible uses of their data both in present and future and be able to give or refuse consent for these potential uses.

The QDR *Guide to Sharing Qualitative Data* (Elman & Kapiszewski 2013, 5–6) further highlights this necessity in the often fast-changing milieu of data sharing:

> Potential study participants must be provided with enough information about the study in which a researcher wishes to involve them and how the information they provide will be used and shared (and anonymized and otherwise protected) that they can make an *informed* decision about whether or not to participate; about how confidential they would like the information they provide to be kept; and about whether and how much of the information they provide will be shared, with whom, and how.

Although, as Bishop (2009) argues, disclosure and informed consent can never truly be full or complete, it is suggested therefore that researchers provide documentation (including plain language statements and consent forms) that give detailed information for the participant on how their data will and may be used in the future. The ANDS provides good guidance for what this documentation and forms should and should not include. Here these recommendations from the ANDS (2018, 7) are quoted in full:

- Consent documentation should:
    - avoid precluding de-identification, publication and sharing of data
    - inform participants how research data will be stored, preserved and used in the long-term
    - inform participants how privacy will be maintained, e.g. by de-identifying data and/or restricting access for secondary use to legitimate researchers
    - state the conditions under which access to the data may be granted to others
    - obtain explicit informed consent for data sharing
    - refer to information provided to participants that describe any risks related to how the data might be used.

- Consent documentation should contain:
    - the level of consent. *The National Statement on Ethical Conduct in Human Research* gives three levels of consent for the future use of data; specific, extended or unspecified (Section 2.2.14). Whichever one is chosen by the researchers must be made clear to the research participants.
    - explicit information on whether the data is to be held in a form which is identifiable, non-identifiable or re-identifiable (for more information see the ANDS Guide on De-Identification).

More detailed information on wording of documentation and content of forms can be found in the full ANDS document. Elman and Kapiszewski (2013, 6) also recommend an information sheet for participants that answers FAQs about data archiving and sharing in plain language such as:

- What is an archive?
- Why put information in an archive?
- How do I know my data will be used ethically?
- What does anonymizing mean?
- How might data be used?
- Who owns the data and what is copyright?
- How do archives store my data safely?

In the process of developing the SOCEY Repository, and the archiving of case studies (see Appendices 1–6), we have observed that forms with multiple levels of consent can be a sound way to allow the participants to best decide how their data may be used both in present and future research. This might include, but is not limited to:

- Permission to archive:

    - personal data (contact details and other sensitive data)
    - the audio and/or video recordings of their interviews
    - other materials they might provide or create in the course of the interview (photographs, written material etc.)

- Permission for (and levels of consent for) sharing:

    - in outputs directly related to the project: journal articles and monographs, conference papers and seminars, project websites
    - with others: including other researchers in the same field, researchers outside the field, government departments working in the field, others outside the academy and/or government that work in the field (for example NGOs working with children and youth

- Permission (and levels of consent) for re-use in the present or future:

    - for related studies in the same field
    - for research outside the field

- Anonymisation of data: whether or not they wish to be identified in outputs and archived research materials.

In some cases, a variety of different levels of consent might be given. For instance, in the *Curriculum Policies Project* archived by Reese (Appendix 4), with regard to consent and confidentiality, the interview consent form gave participants two options:

> 'I agree that comments made in my interview may be quoted and that I may be identified as the source of these, except where I indicate orally during the interview or in subsequent comments on the transcript that I do not wish to be so identified.'
>
> OR
>
> 'I agree that my interview may be drawn on in the overall research project, but it should not be quoted or used in ways that identify me as the source unless I give specific subsequent permission to do so.'

A second question with regards to retention or destruction of materials gave the options of

> 'At the completion of the project (plus five years) I wish my interview tapes and transcripts to be destroyed.'
>
> OR
>
> 'At the completion of the project, I consent to the placing of my interview tapes and transcripts in an archive that may be used by future researchers.'

The consent forms in this case enabled the researcher and archivist to determine whether or not material could be archived and whether it might need to be anonymised.

However, Reese noted that this was not always simple because the lack of a consent form in several cases, the failure of the participant to fill in the form correctly, and the structure of the consent form itself.

> The main challenges regarded interpretation of the consent forms. The consent form text (outlined [in the report]) is vague and badly formatted, and participants' wishes regarding future data usage could not be positively identified in some cases. In these situations it was crucial to err on the side of caution, and to not proceed with archiving unless consent was clearly documented on file.

As a result, of a total of 34 participants, Reese only deposited transcripts for 17 interviews: including seven who did not consent to be archived and eight whose consent forms were not on file. In his summary, he also notes other complexities: including whether or not to archive the consent forms and spreadsheets that mention those that do not wish their material archived.

Reese also noted that:

> A second, related challenge regarded interviewees' consent to be quoted in the study. If a participant gives their consent to be archived, but does not give consent to be 'quoted', does this refer only to publications produced as a direct result of the Curriculum Policies ARC project, or are we to take this as meaning they want their transcript to be anonymised for future archival access? Again, this problem would be obviated by a clearer consent form, with extra questions that cover more specific future situations.

Reese's analysis of this case, fully detailed in Appendix 4, demonstrates the necessity for well-constructed, clear consent forms that are filled out correctly and safely stored during and after the life of the project.

Similar issues were also evident in relation to the *Child Maltreatment and Late Modernity* project. In her summary (Appendix 3), Braithwaite notes that the issue here was possibly related to the number of questions listed on the form, which confused participants and led to sections of the form being skipped, or filling it out incorrectly (e.g., ticking all boxes rather than making a choice). In some cases, they contacted participants to clarify; in others they determined that the conditions to archive had been met regardless of a question being missed or incorrectly completed.

It is not always possible, of course, to cover every potential future usage of research data, particularly if that data is released far into the future, but through documentation such as this, the participant can be given every opportunity to decide on possible uses. If other uses or questions arise, or the researcher (or their designated intermediary) believes that the shared data might be used in a way that the participant might object to, further permissions may be sought wherever possible. This does place an onus on the researcher to ensure that, wherever possible, contact details for the participants are kept up to date, even after the initial lifetime of the project.

The SOCEY pilot also reinforced the importance of seeking consent to archive at the beginning of a project rather than retrospectively. For the project *Globally Mobile Lives and Suburban Government Schooling*, led and archived by Joanne Higginson, retrospective consent was sought from the 15 participants who had not provided consent to archive at the start of the project. In her summary (Appendix 2), Higginson writes that consent to archive is 'best established in the initial relationship building and consent stages of the project', rather than via email some time after the completion of fieldwork.

For the majority of SOCEY pilot projects, consent to archive was sought in the initial stages. However, this was not the case for the project *Our Lives Asylum Seekers* (see Appendix 6). The study is part of a wider longitudinal project. The lead researcher (Jacqueline Laughland-Booÿ) has maintained relationships with the participants and intends to ask for permission to archive at her next round of interviews; however, obtaining this within the timeframes of the SOCEY pilot was not possible. Instead, summaries of the interviews were archived, which consist of a few paragraphs covering the main points touched on in the interviews. The intention is to add to this with the full transcripts in future; however, for now, the summaries provide an indication of the issues covered and the approach taken in the research.

5.2. Anonymisation of data

Decisions surrounding anonymisation of data when archiving, as well as sharing and re-using, need to be considered carefully. While anonymisation of data in both quantitative and qualitative research outputs is common practice in the social sciences in order to maintain the confidentiality of the participant, the researcher should consider whether this is appropriate for the long-term archiving, sharing and re-use of that data. Does excising identifying details, including real names, locations, names of businesses or localities etc. serves to de-contextualise the data for usages beyond the study for which it was initially collected? As Corti et al. (2000) argue, 'the key issue … is that it is important to arrive at an appropriate level of anonymisation to ensure that data are not distorted to a degree which lessens their potential for reuse' (see also ADA 2014).

From a historical perspective, what value might be contained in non-anonymised data for future researchers? The case of historic census data in the United Kingdom, for instance, when compared with Australia is telling. The preservation of detailed nineteenth-century census data in the UK provides researchers from various disciplinary fields with the ability to make comparisons beyond the basic qualitative demographic data available in the Australian context, where the

original returns have been destroyed. Likewise, they have proved invaluable to descendants, with the growing interest in family history (see for example Corti 2018 on the usefulness of non-anonymised data to historians).

Questions of anonymity and context, while not irrelevant, are of a different order when the interviewee's identity and location is key to the significance of the history. Sharing current or recent qualitative interviews arouses different concerns, and rightfully so, in part because the participants enter the encounter with different purposes and expectations, and research aims can be so diverse. Consideration of course needs to be given to the effects of open access for interview transcripts (on children, communities) and this necessarily offsets the opportunism of re-use.

From the perspective of qualitative data, anonymisation of data can be problematic. As Bishop (2009) discusses, it requires excellent knowledge of the data itself in order to fully anonymise or de-identify but without distorting the data for both present and future use. As Corti et al. (2000) argue, facts or stories about the participant might still allow them to be identified. In the field of educational research, for example, while changing the name of a school or not referring to it by name might provide some anonymisation, often the contextual data that must be provided in order to fully analyse the material will render that school (and potentially individuals) identifiable (see also Fitzgerald, Appendix 1).

In one of the projects being archived in the SOCEY Repository, *Making Futures*, these questions have been paramount. In the initial planning stages, the lead researcher gained permission from all three schools involved in the study for their institution to be named explicitly in outputs from the study. This was because the study was concerned with understanding how students' experiences of schooling were embedded in the history of the school's locality. In archiving the project, Popovski changed only the names of places and locales if the identity of the participant could be traced back to them (for example the name of a small business, or the street on which it was located) (see Appendix 5). Additionally, student participants all gave consent at the start of the project for their real names to be used on the website and for any transcribed material to be made available on the website and archived. However, in the course of undertaking the longitudinal study, the research team decided to defer disclosure of full transcripts and names. This was in recognition of the age of participants (14–19), concerns about exposure of personal details, and some questions as to whether the scope of their consent was in fact fully appreciated by or their parents/guardians at the time the formal consent process was negotiated. There were also methodological considerations, in that full access to transcripts would also have been available to participants and could well influence what they then felt able to say in subsequent interviews, when the aim of the study was not to foster this, but to allow participants to begin each round of interviews afresh, not in the explicit shadow of previous interviews – these are matters amplified in longitudinal research.

With informed consent of participants, non-anonymised data might be legally shared, but the researcher must consider the ethical dimensions and continue to revisit these throughout and after the life of the project. With detailed information provided as to how their data might be used, as discussed in the preceding section, it is often the case that a research participant will gladly allow their data to be archived, shared and re-used without being fully anonymised. Inevitably, within the boundaries of ethics, each researcher must make the decision whether and how much to anonymise qualitative data intended for archiving, sharing and re-use.

This also holds for retrospective deposit of research data into repositories where the participant may not have been approached for specific permission to archive. The ADA's *Archiving Qualitative Data from a Completed Project* (2014) suggests that if documentation and consent

forms specifically refer to the destruction of all or part of the data, that this data cannot be archived. However,

> if specific consent to archive participants' data has not been obtained, the ethics of archiving the data must be carefully considered. ADA advises that consent gained to use data for 'research purposes only' would be considered sufficient to allow for archiving. This allowance is in accordance to the 'Information Privacy Principles' in the *Privacy Act*, *1988*.

In the case of the *Schooling Memories: Educating the Adolescent* project, which ran from 2009 to 2013, the 77 participants were given the choice to have their data archived or destroyed but were 'also given an option to remain anonymous'. Of the 77, only ten participants selected the option of 'I agree that my interview may be drawn upon in the overall research project, but it should not be used in ways that identify me as the source unless I give subsequent permission to do so.' As a result, the interviews of these participants were anonymised and their original transcripts not included in the archive. Fitzgerald noted in her report (see Appendix 1) that the researchers' forward planning with regard to permissions was advantageous almost five years later when it came to making decisions about archiving the project:

> By including the creation of an oral history archive as a goal from the beginning of the project, and therefore being able to explicitly ask the question in the consent form, the investigators in this project have been able to avoid ethical concerns about the archiving of the data. It has also meant that they have been able to create an archive where the bulk of the transcripts available have not been de-identified, and therefore create a richer and more valuable source of information for future researchers, as they will be able to put the responses provided into greater context. This has the flow on effect of reducing the risk of incorrect conclusions being drawn by future researchers relying on incomplete information.

Fitzgerald does comment though that there are still questions surrounding the inclusion of the original non-anonymised transcripts of the ten de-identified participants in the archive – while the non-anonymised materials are important for record-keeping and the researcher, a decision still needs to be made of a) whether to include them at all in the archive and b) present and future access levels (see Section 5.3 regarding access levels).

In developing the SOCEY Repository alongside the ADA, our pilot research project deposits to the archive were given options for anonymising data by referring to guidelines from the ADA (ADA 2014) – *Archiving Qualitative Data from a Completed Project* – along with those from several other data archives, including UK Data Service (UKDS 2012–2019b) and the QDR (2013), on which the ADA have based their own guidelines.[5]

There is often a variety of data that may need to be anonymised: including base data such as names, addresses, email, phone number; text-based data (including transcripts and other written materials); and audio and video recordings. As per the QDR guidelines, reproduced here in full, based on the UKDA guidelines, this includes:

- Removing major (direct) identifying details (e.g., real names, locations); replacing them with pseudonyms, replacement terms (e.g., "paternal grandfather"), vaguer descriptors or some coding system (where appropriate) consistently throughout the project; and devising and using a cross-referencing system for pseudonyms that will not be made available to users;

  [ADA (2014) guidelines stipulate that 'Particular types of information, such as names, residential address, or workplace address, cannot be archived, unless the participant has given express consent for the researcher to do so. This is ADA policy, in accordance with the *Privacy Act 1988*.']

[5] ANDS (2018) also has a brief guide to anonymising data but refers back to other archives' guidelines for particulars and does not delve into detail from a specifically qualitative data perspective.

- Removing information in a transcript or notes from a human encounter that may reveal the identity of project participants;
- Aggregating or reducing the precision of information or a variable, e.g., replacing date of birth by age groups or city names by province names;
- Generalizing the meaning of detailed text, e.g., replacing a doctor's detailed area of medical expertise with an area of medical specialty;
- Restricting the upper or lower ranges of a variable to hide outliers;
- Noting the replacement of identifying details in text and the removal or modification of information in a meaningful way (for instance, in transcribed interviews, indicating replaced text with [brackets] or using XML markup tags <anon>.....</anon>);
- Creating an anonymization log (stored separately from the anonymized data files) of all replacements, aggregations, or removals.
(QDR 2013)

The ADA notes 'that these should only be done on a copy of the file and not the master copy which remains unedited', with any changes made in the separate log file.

With regard to audio and video research materials:

> Digital editing can be undertaken to remove identifying detail by bleeping out names, altering the pitch of audio recordings, and pixelating sections of video images to disguise participants, sensitive material or other identifying images. Such practices can significantly reduce the useability of the data though, and can be time consuming and expensive. (ADA 2014)

As with other materials, researchers should take care to include permission to archive and not anonymise audiovisual materials in their consent forms. If this has not been given the material should be digitally edited, retrospective permission gained for archiving, or impose access restrictions on the material to allow it to be archived. For the SOCEY projects, audio files were not able to be de-identified and so were not archived with the datasets. In her summary of the *Schooling Memories* project (Appendix 1) Fitzgerald writes:

> If it is determined that these should also be included, serious consideration will need to be given to if the audio files of the de-identified interviews could be included. The nature of the de-identification process was such that there were redactions on nearly every, if not every, page of the transcripts. While the redacted transcripts could be used as a guide, the process of de-identifying the audio files would be onerous and potentially make the files unusable.

UKDS (2012–2019b) also notes the importance of forward planning in preparing research data:

> Consideration should be given to the level of anonymity required to meet the needs agreed during the informed consent process. Pre-planning and agreeing with participants during the consent process, on what may and may not be recorded or transcribed, can be a much more effective way of creating data that accurately represent the research process and the contribution of participants.

Usually decisions about what should or should not be anonymised and to what extent need to be made on a case-by-case basis (ADA 2014). Each of the SOCEY pilot archivers made decisions about anonymisation based on the particular purposes and permissions of their individual projects (see Appendices 1–6). Their decisions differed across a number of areas, including decisions to anonymise or identify place names and remove sensitive material or references to difficult events such as illness. Archivers were required to think about both the extent to which participants could potentially be identified by information contained within the transcript, and the importance of ensuring the transcript would remain a useful document for future researchers. In Appendix 1 Fitzgerald writes:

> The process of de-identification is one of constant evaluation. Even when I had established a set of protocols for myself, these were constantly being challenged with new or slightly different circumstances. While retaining anonymity of the interview subjects (as far as possible) was central to this decision making process, as can be seen from the examples given above, there was constant consideration regarding the need to de-identify something, and, when it was a grey area, balancing the value of the information against the possible risk of identification.

Automated Data Anonymisation

As noted by the ADA (2014), QDR (2013), UKDS (2012–2019b), and others, anonymisation or de-identification of data can be time consuming and expensive, even when planned ahead of time. For the SOCEY pilot, the costs associated with archiving each project ranged from AUD$3000 to $5000 per project, or between 60 to 100 hours of research assistant support, with anonymisation taking up the majority of time in most cases.

As one possible approach to make anonymisation at scale more viable, we investigated the current state of software-based anonymisation available for qualitative data in the social sciences.

Four software packages were selected and investigated with a sample interview transcript to see the efficacy of the software. These packages are all freely available on the internet (three are open-source, one is closed-source). The software packages were:

1. UK Data Service Text Anonymiser (UKDS 2012–2019d).
2. The Irish Qualitative Data Archive provides an open-source anonymisation software written in Java (keithodulaigh 2013; Maynooth University n.d.)
3. NLM-Scrubber, a medical text anonymisation software by the U.S. National Library of Medicine (USNLM 2019)
4. Netanos, a named entity recognition based anonymiser (Netanos 2018).

All four software packages both (1) identify terms that need to be replaced with pseudonyms; and (2) replace those terms with the appropriate pseudonyms; however, they are substantially different in the extent and accuracy of the approach. A full account of the individual approaches taken, and the extent of their efficacy is provided in Appendix 7 (written by Geordie Zhang).

The Netanos software, which uses natural language processing (NLP) based term identification, followed by a software automated replacement of all identified terms with their pseudonyms, is considered the most promising. However, NLP can only do entity-based identification of terms to be anonymised (such as persons, organisations, locations), cannot address some of the subtler nuances of term identification, and is not perfect in its identification of terms based on entities.

The software overall offer potential to reduce workloads with regard to anonymisation of qualitative research data but do not replace the value of a researcher with detailed knowledge of the project making decisions about anonymisation in context.

5.3. Setting Access Conditions[6]

Setting access conditions is another recognised method of protecting an individual's confidentiality while at the same time allowing for the sharing and re-use of research data, whether anonymised or not. The UKDS (2012–2019b) says of access conditions:

[6] Significant content for this section was provided by Janet McDougall.

Sensitive and confidential data can be safeguarded by regulating or restricting access to and use of the data. Access controls should always be proportionate to the kind of data and level of confidentiality involved.

When regulating access, consider who would be able to access your data, what they are able to do with it, whether any specific use restrictions are required, and for how long you want the data to be available.

The ADA provides varying levels of access to data from open and shareable, to highly mediated, including secure access only. The conditions under which the data may be made available to other researchers are determined by the depositor of the data. Depositors are able to choose whether they share all or parts of the data, as long as there is utility in the dataset deposited for sharing. The data deposit can be a subset of the full dataset if it is not possible to make the entire project data available.

Data is made available under five broad categories of access, summarised as:

- Open Access: There are no restrictions on access. The user will be required to adhere to and accept the Terms and Conditions of Use and any copyright restrictions laid out by the Data Owner.
- Recorded Open Access: There are no restrictions on access. The user will be required to adhere to and accept the Terms and Conditions of Use and any copyright restrictions laid out by the Data Owner. The user will also be required to complete an Access Guestbook for the collection of basic user information prior to download.
- ADA Managed Access: Access approval will be undertaken by ADA on behalf of the Data Owner. The user will be required to hold an ADA Dataverse Account and to complete an Access Guestbook as part of the access request process through Dataverse. Prior to downloading any material, the user is to agree to, and accept the Terms and Conditions of Use and any copyright restrictions laid out by the Data Owner.
- ADA Facilitated Access: The ADA will facilitate the access request through the Data Owner or an authorised representative, and prior to downloading material, the user is to agree to and accept the Terms and Conditions of Use and any copyright restrictions laid out by the depositor.
- Non-Standard Access: The Data Owner is required to liaise with the ADA Director in order to agree the details surrounding access to the dataset.

Depositors can also impose embargo periods on their full data or subsets thereof, whereby no access to the data would be permitted until after the date specified. This can assist in ensuring that other researchers do not pre-empt findings prior to publication. If there are protection concerns regarding participants conditions can be set to mediate these issues. At the end of the embargo period, the data may be released under the access conditions set for the data. Catalogue information and metadata about a study and its contents is also not subject to conditions but is made freely accessible to enable researchers to locate relevant data.

Users are required to acknowledge that materials will only be used to produce information of an analytical nature and cannot be used for commercial or financial gain, or for non-analytical purposes, without the express written permission of the ADA, as outlined in the ADA Terms and Conditions of Use, for ADA Managed, Facilitated and Non-Standard Access.

The latter includes allowing access to secondary parties or attempting match data with other information for the purposes of attempting to identify individuals. In disseminating any outputs obtained from analysis of the data, users also agree to acknowledge the original depositors and the ADA and declare that those who carried out the original analysis and collection of the data bear no responsibility for the further analysis or interpretation of it.

For ADA Managed, Facilitated and Non-Standard datasets, supplementary materials (such as project descriptions and blank consent forms) are made available for download, but the project data are not able to be accessed without permission from the ADA or the lead researcher.

It is important that shared research materials are discoverable and useable by others. The European Commission Directorate-General for Research and Innovation's (ECDGRI) *Guidelines on FAIR Data Management in Horizon 2020* (2016) emphasise that data should be FAIR, meaning Findable, Accessible, Interoperable and Reusable. ANDS (n.d.) defines the four principles as such:

> Findable: This includes assigning a persistent identifier (like a DOI or Handle), having rich metadata to describe the data and making sure it is findable through disciplinary discovery portals (local and international).
>
> Accessible: This may include making the data open using a standardised protocol. However the data does not necessarily have to be open. There are sometimes good reasons why data cannot be made open, for example privacy concerns, national security or commercial interests. If it is not open there should be clarity and transparency around the conditions governing access and reuse.
>
> Interoperable: To be interoperable the data will need to use community agreed formats, language and vocabularies. The metadata will also need to use a community agreed standards and vocabularies, and contain links to related information using identifiers.
>
> Reusable: Reusable data should maintain its initial richness. For example, it should not be diminished for the purpose of explaining the findings in one particular publication. It needs a clear machine readable licence and provenance information on how the data was formed. It should also have discipline-specific data and metadata standards to give it rich contextual information that will allow for reuse.

While FAIR principles were largely developed with quantitative data in mind, they are to a great extent able to be applied to the practice of archiving, sharing and re-using qualitative data. By and large, these principles under pin the way that data is deposited in the ADA in general and will be in the SOCEY Repository specifically.

The data deposited in the SOCEY Repository receives a DOI once published and providing rich metadata is encouraged, in order to make the data *findable*. This findability of data by providing rich metadata was a key concern of researchers in both the 2018 seminar on data sharing held by the SOCEY team and the recent 2019 SOCEY Roundtable.

As discussed above, archiving qualitative data in the SOCEY Repository makes research data in the field *accessible*. Although open access may not be appropriate for all data deposited by researchers, they are able to choose levels of access and periods of embargo, as discussed above.

Likewise, much of the qualitative data deposited in the SOCEY Repository may be able to be *reused* by other researchers in the present or future, depending again on the appropriateness of that usage. Such reuse, as indicated above, can be mediated either by the researcher or the ADA adhering to conditions set by that researcher.

Interoperability may be less applicable to qualitative data in some ways, as often there are not necessarily 'community agreed formats, language and vocabularies' (ANDS n.d.) for the diverse types of qualitative data that might be included in such an archive. Most quantitative datasets tend to be structured with a set format and sequence, whereas most interview transcripts/free text tend to be unstructured or semi-structured. This renders the meaning of interoperable less clear with regard to qualitative data. One of the main values of qualitative research is that the researchers are free to explore issues with the interviewee without being restricted to a fixed instrument or strict

interview schedule. This can make interoperability (in the traditional quantitative sense) harder to achieve, because it is more likely that each qualitative dataset will be different at the meta level. Interoperability for qualitative data therefore is more applicable at the level of the metadata rather than the data itself; at the level of the systems used (for example, requiring all interviews to be saved in UTF-8 character encoding); and, if natural language processing (NLP) does become something of interest for qualitative data repositories (e.g., NLP-based annotation of interview transcripts), interoperable could mean having standards on the structure of the storage of NLP training and validation datasets in the context of the repository (e.g., human-annotated interview transcripts used for the NLP models). From this perspective, interoperability is accomplished through the other principles of the FAIR model as adopted by the ADA in its standards.

For the SOCEY pilot, all projects were made available under the ADA Facilitated Access category to allow the lead researcher to maintain control of access to the dataset. ADA Facilitated Access is required for sensitive data and it was determined that the qualitative data being archived in the SOCEY repository was sensitive even where anonymised given the challenges of completely anonymising qualitative data discussed above. An additional embargo was also placed upon a number of datasets to allow the lead researcher to complete their own analysis prior to making the materials available. Restricted access is also useful for those researchers that simply wish to utilise a digital repository to store their research data, rather than share it.

The ADA Facilitated Access category was also seen as important for ensuring that the research is used appropriately in future. At both the 2018 seminar on data sharing and the recent 2019 SOCEY Roundtable, attendees raised concerns with research on children and young people being used inappropriately by media organisations, particularly where interviews contained sensitive material. In her summary (Appendix 2), Higginson also notes that, in her view, those that provided consent to archive 'were motivated by ideas of historicity and were reassured that their transcripts would only be accessed for legitimate research purposes only – i.e., they cannot be uncovered via a simple internet search'.

5.4. Selecting, preparing and formatting materials for deposit

In the field of qualitative data, there can be a long list of potential materials that might be selected for archiving. In many cases in studies of education, youth and childhood, these might often be audiovisual interviews, interview transcripts, photographs of participants, drawings or written material provided by participants, consent forms, plain language statements, spreadsheets of data, coded research materials, summaries of interviews, and more. The QDR *Guide to Sharing Qualitative Data* (2013) provides detailed list of potential research material that might be considered for archiving and is a useful resource for researchers to begin to consider what they might archive.

Selecting materials is not necessarily a straightforward process. In her summary of the *Childhood Maltreatment and Public Inquiries* project (Appendix 3), Braithwaite describes the question of what to include as 'the biggest question we had' in preparing the dataset. Along with the project lead researcher, she writes that they undertook a process of interrogating what could be deposited, asking what value it had and the ethical considerations of including it in the dataset. She writes: 'A major consideration was consistency across the data-set – we wanted to deposit a consistent set of data which made the archive clear to use and understand.'

The development of a collection (or selection) policy is crucial for the researcher in allowing them to decide what to archive in a qualitative data repository; clearly not all potentially relevant material can be or perhaps should be collected. The processes in place for making decisions about what is and is not to be collected and then what is and is not made open, or available in stages and

sequences could offer guidance for data sharing among qualitative researchers working in the academies.

Shared data need to be understandable and interpretable to scholars beyond those who collected or generated them. So that those subsequent consumers of shared data can make informed and effective use of them, the data must be accompanied by documentation that describes the project to which the data are connected, the data themselves, and the processes by which they were collected or generated. In terms of qualitative research therefore, it is essential that contextual materials are also included in the research project archive. This might include, but are not limited to, research notes, summaries, research outputs (including journal articles), and project website content. The UK Data Service (2012–2019c) writes:

> A crucial part of ensuring that research data can be used, shared and reused by a wide-range of researchers, for a variety of purposes, is by taking care that those data are accessible, understandable and usable. This requires clear data description, annotation, contextual information and documentation that explains how data were created or digitised, what data mean, what their content and structure are, and any manipulations that may have taken place.
>
> Creating comprehensive data documentation is easiest when begun at the onset of a project and continued throughout the research.

For the SOCEY pilot, supplementary materials archived included ethics applications, blank plain language statements and consent forms, project descriptions, website pages, interview schedules, anonymisation process summaries and summaries of interview transcripts. These materials provide a comprehensive overview of the context of the research and the process through which data was collected. As noted above, they are not subject to the same restrictions as the primary data materials but are made freely available for download.

Such materials provide value for re-use of the datasets but also important benefits in terms of deepening understanding of the methods and nature of qualitative research. In her summary of her archiving experiences (see Appendix 5), Popovski writes,

> As a graduate research student with no prior knowledge of the research project, one unexpected benefit of having access to research materials and interview data in the capacity I did, was exposure to the interview skills and techniques employed by the researcher. This was beneficial as it gave me insight into how to conduct interviews in an engaging and meaningful manner.

These comments highlight the value of data archiving and sharing for research training, and their potential use in qualitative methods teaching. They point to new opportunities in enabling graduate researchers to gain in-depth, in-context knowledge of the practices and methods of qualitative research, which is often difficult to understand in the abstract.

Selecting files for archiving but not sharing is also an option available to researchers. For the *Making Futures* and *Schooling Memories* projects, original transcript files and consent forms were also uploaded as ‘for archiving only’, meaning that these materials are not included as part of the published dataset and not visible to other users. This was chosen to allow for safe archiving of the research materials whilst ensuring those materials are not shared with others.

6. KEY LEARNINGS FROM THE SOCEY PILOT

Based on the above discussion and our SOCEY pilot experiences, we propose the following principles to guide further work in this area:

1. Decisions regarding data archiving and sharing should ideally be considered from the start of the research process, and further work is needed to encourage researchers to do this.

   The challenges we faced in recruiting projects for the SOCEY pilot were, in part, a result of the limited number of projects which have been completed with approvals for archiving and data sharing in place. Further work is needed to encourage greater consideration of decisions about archiving and sharing in qualitative research with children and young people, including via ethics approvals and grant applications processes.

2. Standardised wording is needed for developing consent forms which provide appropriate and unambiguous archiving options

   The experiences of the SOCEY pilot highlight some of the problems raised where consent forms are not written clearly or are inadvertently ambiguous. It is recommended that standardised language on consenting to data archiving and sharing be developed which includes options for identification/deidentification as well as consent to archive.

3. Protocols can be established for best practice in anonymisation of research materials, but this process needs to take into account the particular purposes and contexts of the research project

   Automated data anonymisation software could offer potential workload benefits but is unlikely to replace the value of having a researcher making decisions about anonymisation in the context of particular projects. The SOCEY pilot projects highlight the importance of allowing for difference in how anonymisation is conducted, while also offering some potential model approaches for other projects to consider.

4. Ethical issues relating to consent and identification should be managed at the outset of the project, but also revisited throughout the research and archival process.

   The ethical issues raised by data sharing will be different for each project. For qualitative sociological research, it is important that researchers address ethical concerns at each stage of the project, from design onwards. While there are strong and sometimes compelling calls for openness, attention to limits and to what should be and should not be shared is crucial to the conversation, and this is particularly the case with research involving children and young people.

5. Archived qualitative datasets have value not just in terms of the potential for re-use but also in terms of deepening understanding of the methods and nature of qualitative research, and the selection of materials for deposit should take this into consideration

   The material archived in the SOCEY pilot provides a valuable resource for understanding in detail the methods and nature of qualitative research, from the drafting of ethics applications

and interview questions to forms of qualitative interviewing. In doing so, it offers resources of use to qualitative methods teachers and early stage or emerging researchers as well as to potential end-users in better understanding the different ways qualitative research works in practice. Decisions about the selection of materials should take this into account and ideally include a broad range of supplementary material providing a detailed account of how the research was designed and the considerations and practices that were part of this.

6. Qualitative research involving children and young people should generally be considered to contain sensitive data, and be made available via facilitated access, unless this is determined as not necessary by the lead researcher.

   Complete anonymisation of qualitative interview data is likely impossible as there always remains a possibility that a participant's identity could be guessed. As such, we choose to archive all SOCEY pilot projects under facilitated access, and to recommend that this approach be taken for other similar projects in future.

7. It is essential that those who have created and curated the data initially maintain full control over how data is managed, mediated, shared and accessed.

   Decisions about data archiving and sharing need to remain the purview of the appropriate researchers, and not overly determined by repository processes. This is particularly the case for qualitative research involving children and young people where concerns about processes, methodologies and ethics may prevent researchers from depositing and sharing their data.

7. NEXT STEPS

Building on the models available from other repositories, we plan to develop specific guidelines for the SOCEY Repository as part of creating documentation that is relevant to and supports qualitative data deposit, archiving and re-use for researchers working in Australian research and regulatory environments.

These guidelines will provide researchers with a step-by-step process for organising their qualitative data for deposit, prompting them to consider what they do and do not wish to archive, and which data they wish to have as shareable and re-usable by current or future researchers.

The repository will also be opened up for interested parties to deposit data after these guidelines are complete. Further, the SOCEY community of practice will continue to grow, encompassing new projects from both inside and outside the academy.

With interest in making research data more accessible from numerous quarters, creating a community of practice in the field of childhood, education and youth studies, as well as providing a supported space for researchers to discuss, access and deposit their data, represents important and promising directions for qualitative researchers across the social sciences.

BIBLIOGRAPHY

Andrews, M. 2011. Never the Last Word: Revisiting Data 1. In Molly Andrews, Corinne Squire & Maria Tamboukou (eds.) *Doing Narrative Research*. London: Sage. https://dx.doi.org/10.4135/9780857024992

ARC 2017. ARC Open Access Policy Version 2017.1 https://www.arc.gov.au/policies-strategies/policy/arc-open-access-policy-version-20171

ARC 2019 (last updated 27/09/19) Research Data Management https://www.arc.gov.au/policies-strategies/strategy/research-data-management

Australian Data Archive (ADA) 2014. *Archiving Qualitative Data from a Completed Project*. Canberra: Australian Data Archive

Australian Data Archive (ADA) 2018. *Depositing Data*. Australian Data Archive https://ada.edu.au/depositing-data/

Australian Data Archive (ADA) 2019. *ADA Self Deposit Wiki Guide* https://docs.ada.edu.au/index.php/Main_Page

Australian Data Archive (ADA) n.d. https://ada.edu.au

Australian National Data Service (ANDS) 2018. *De-identification* https://www.ands.org.au/__data/assets/pdf_file/0003/737211/De-identification.pdf

Australian National Data Service (ANDS) n.d. *The FAIR Data Principles*. https://www.ands.org.au/working-with-data/fairdata

Bishop, L. 2005. Protecting Respondents and Enabling Data Sharing: Reply to Parry and Mauthner, *Sociology* 39.2, 333–336.

Bishop, L. 2009. Ethical Sharing and Reuse of Qualitative Data. *Australian Journal of Social Issues* 44.3, 255–272. https://doi.org/10.1002/j.1839-4655.2009.tb00145.x

Bornat, J. 2003. A Second Take: Revisiting Interviews with a Different Purpose. *Oral History* 31.1, 47–53.

Bornat, J. 2013. Secondary Analysis in Reflection: Some Experiences of Re-use from an Oral History Perspective. *Families, Relationships and Societies* 2.2, 309–317.

Broom, A., Cheshire, L. and Emmison, M. 2009. Qualitative Researchers' Understandings of Their Practice and the Implications for Data Archiving and Sharing. *Sociology* 43.6, 1163–1180. DOI: 10.1177/0038038509345704

Callison, C., Roy, L. & LeCheminant, G.A. (eds.) (2016). *Indigenous Notions of Ownership and Libraries, Archives and Museums*. Berlin & Boston: Walter de Gruyter.

Cheshire, L., Broom, A. & Ellison, M. 2009. Archiving Qualitative Data in Australia: An Introduction. *Australian Journal of Social Issues* 3, 239–254.

Christou, T. 2009. Gone but not forgotten: the decline of history as an educational foundation. Journal of Curri*c*ulum Studies, 41:5, 569-583.

Corti, L., Day, A. & Backhouse, G. 2000. Confidentiality and Informed Consent: Issues for Consideration in the Preservation of and Provision of Access to Qualitative Data Archives. *FORUM: Qualitative Social Research* Sozialforschung 1.3.

Corti, L., Van den Eynden, V., Bishop, L. & Woollard, M., 2014. *Managing and sharing research data: a guide to good practice*. London: Sage Publications.

Corti, L. 2018. Data Collection in Secondary Analysis. In Uwe Flick, *The SAGE Handbook of Qualitative Data Collection*. London: Sage Publications.

Crossen-White, H. 2015. Using digital archives in historical research: What are the ethical concerns for a 'forgotten' individual? *Research Ethics* 11.2, 108–119.

Daniels, D., Senior, K. & Edmonds, F. 2019. Pathways for Indigenising the Archive: A Ngukurr community collaboration. Seminar Paper, Faculty of Arts Digital Studio. University of Melbourne 8 October.

Dataverse n.d. *The Dataverse Project* https://dataverse.org/

Dening, G. 1998. Past Imperfect. *The Australian Review of Books* 3.3, 4–5.

Elman, C. & Kapiszewski, D. 2013. *A Guide to Sharing Qualitative Data: Version 1.3 for Pilot Projects*. Syracuse: Qualitative Data Repository (QDR), Center for Qualitative and Multi Method Inquiry (CQMI).

European Commission Directorate-General for Research and Innovation (ECDGRI) 2016. *Guidelines on FAIR Data Management in Horizon 2020*. https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

Fielding, N. 2004. Getting the Most from Archived Qualitative Data. *International Journal of Social Research Methodology* 7, 97–104.

Grant, R. 2015. Identifying HSS Research Data for Preservation: A Snapshot of Current Policy and Guidelines. *New Review of Information Networking* 20:1-2, 97-103, DOI: 10.1080/13614576.2015.1110400

Goodman, J. & Grosvenor, I. 2009. Educational research – history of education a curious case? *Oxford Review of* Education 35(5), 601-16.

Hammersley, M. 1997. Qualitative data archiving: some reflections on its prospects and principles. *Sociology* 31.1, 131-141.

Hammersley, M. 2010. Can we re-use qualitative data via secondary analysis? Notes on some terminological and substantive Issues. *Sociological Research Online* 15:1, 5.

Heaton, J. 2004. *Reworking Qualitative Data*. London: Sage Publications.

keithodulaigh 2013. *IQDA Qualitative Data Anonymizer*. https://sourceforge.net/projects/datatool/

Klein, L.F, & Gold, M.K. 2016. Digital Humanities: The Expanded Field. In Lauren Klein & Matthew Gold (eds.), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press.

Making Futures n.d. http://makingfutures.net

Mauthner, N., Parry, O., & Backett-Milburn, K. 1998. The Data are Out there, or are They? Implications for Archiving and Revisiting Qualitative Data. *Sociology* 32.4, 733–745.

Maynooth University: National University of Ireland, Maynooth n.d. *Irish Qualitative Data Archive* https://www.maynoothuniversity.ie/iqda

McLeod, J., O'Connor, K. & Davis, N. 2018. Data Sharing and the Challenges Facing Educational Researchers. In *EduMatters Research*, Australian Association for Research in Education. Available at https://www.aare.edu.au/blog/?p=3214

Melbourne School of Population & Global Health n.d. *Indigenous Data Network* https://mspgh.unimelb.edu.au/centres-institutes/centre-for-health-equity/research-group/indigenous-data-network

Moore, N. 2007. (Re)using Qualitative Data?. *Sociological Research Online* 12:3, 1.

Netanos 2018. *Netanos* http://netanos.io/

NHMRC 2018. Australian Code for the Responsible Conduct of Research, 2018. https://www.nhmrc.gov.au/about-us/publications/australian-code-responsible-conduct-research-2018

O'Connor, K. 2019. Workshop: Engaging with Qualitative Data in Studies of Childhood, Education and Youth, *SOCEY*. https://www.socey.net/2019/10/04/workshop-engaging-with-qualitative-data-in-studies-of-childhood-education-and-youth/

Parry, O. & Mauthner, N. 2004. Whose Data Are They Anyway? Practical, Legal and Ethical Issues in Archiving Qualitative Research Data. *Sociology* 38.1, 139–152.

Qualitative Data Repository n.d. https://qdr.syr.edu

Reanimating Data n.d. http://reanimatingdata.co.uk/

Savage, M. 2010. *Identities and Social Change in Britain since 1940.* Oxford: Oxford University Press.

Skrbiš, Z., Tranter, B., Parsell, C., Smith, J., & Laughland-Booÿ, J. 2016. *Our Lives: The First Ten Years 2006–2016*. Melbourne: Social Futures and Life Pathways of Young People in Queensland.

Smith, E. 2008. Using Secondary Data in Education and Social Research. Maidenhead, UK: Open University Press, McGraw-Hill.

SOCEY 2019. https://socey.net

Thorpe, K. & Galassi, M. (eds.) 2014. Rediscovering Indigenous Languages: The Role and Impact of Libraries and Archives in Cultural Revitalisation. *Australian Academic and Research Libraries* 45.2, 81–100.

Thor Tureby, M. 2013. To Hear with the Collection: The Contextualisation and Recontextualisation of Archived Interviews. *Oral History* 41.2, 63–74.

UK Data Archive (UKDA) 2014. *Qualitative Data Collection Ingest Processing Procedures*. Colchester: UK Data Archive.

UK Data Service (UKDS) 2012–2019a. *Qualitative and Mixed Methods Data*. UK Data Service https://www.ukdataservice.ac.uk/get-data/key-data/qualitative-and-mixed-methods-data

UK Data Service (UKDS) 2012–2019b. *Legal and Ethical*. UK Data Service https://www.ukdataservice.ac.uk/manage-data/legal-ethical/

UK Data Service (UKDS) 2012–2019c. *Document Your Data*. UK Data Service https://www.ukdataservice.ac.uk/manage-data/document

UK Data Service (UKDS) 2012–2019d. *Anonymisation: Qualitative* https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative.aspx

University of Sheffield n.d. Research Data Repositories. University of Sheffield Library. https://www.sheffield.ac.uk/library/rdm/repositories

US National Library of Medicine (USNLM) 2019. *NLM-Scrubber* https://scrubber.nlm.nih.gov/ last updated 20 May 2019.

APPENDICES

Appendix 1: Schooling Memories: Educating the Adolescent, 1930s–1970s

Project led by Julie McLeod and Katie Wright. Summary prepared by Emily Fitzgerald

Project summary

Schooling Memories comprises the oral history component of *Educating the Adolescent: An Historical Study Of Curriculum, Counselling and Citizenship in Australia 1930s–70s,* an ARC-funded research project undertaken by Professor Julie McLeod (Melbourne Graduate School of Education) and Associate Professor Katie Wright (La Trobe University). Running from 2009 to 2013, the project used both documentary research and oral history research to undertake a cultural history of educational reforms and ideas of how Australian adolescents could be educated for future citizenship. The project particularly targeted three decades of educational upheaval, being the 1930s, 1950s, and 1970s.

As part of this research, the investigators interviewed students, teachers, guidance councillors and school psychologists, and curriculum and policy personnel from these key periods. They aimed to gain recollections of participants' experience of school as a student and (where relevant) as a teacher or other education professional, the purpose of school, the role of young people, and the priorities in curriculum and in counselling. Ultimately, the investigators interviewed 84 people across 71 interviews.

One of the intended project outcomes was the creation of an oral history archive that would be available for future researchers.

Materials archived

This oral history archive is primarily made up of transcripts from the oral history interviews undertaken as part of this project, this being 59 individual interviews and six group interviews.[7]

As it was intended from the beginning of the project that an oral history archive would be created, the investigators were able to include the question in the consent form for all interview participants. Interview subjects had the option to choose between 'At the completion of the project, I consent to the placing of an audio recording and transcription of my interview in an archive that may be used by future researchers,' or 'At the completion of the project (plus five years) I wish that the audio of my interview and transcripts be destroyed'. Consent was provided by 77 participants for their interviews to be included.

Participants were also given an option to remain anonymous, with ten participants selecting the option of 'I agree that my interview may be drawn upon in the overall research project, but it should not be used in ways that identify me as the source unless I give subsequent permission to

[7] Only five of the six group interviews are listed as such in the documentation. The sixth group interview was a husband and wife interviewed together, so only one transcript was produced, though the interview subjects each signed an individual consent form. For further information on this interview, see Case Study 10.

do so'. For these participants, both a de-identified transcript and the original transcript will be archived, with the original transcript not being made available to researchers.

In addition, supplementary materials such as blank consent forms, plain language statements, website pages, and a broad guide of the process undertaken while de-identifying transcripts have also been included in the archive materials.

Archival process

The interviews were conducted by Julie McLeod and Katie Wright, and transcripts prepared during the project. Preparation of transcripts had also included participants having the option to review transcripts, and edits being made and reviewed as part of that process. Forty-four of the transcripts included in the archive were edited.

Summaries were also created for oral history transcripts, with 35 of the included transcripts having summaries made.

I was engaged to prepare the materials for archiving. This work has included:

- Checking consent forms to ensure they matched the consent levels recorded;
- Briefly checking each transcripts to ensure edits completed, correcting minor errors found, and recording the interview date;
- Identifying which of the transcripts to be de-identified had had pseudonyms created;
- De-identifying the ten transcripts where permission to be identified had been withheld, keeping a record of all amendments made;
- Liaising with the Australian Data Archive (ADA);
- Entering metadata and data description into the data repository shell created;
- Formatting and uploading the transcripts to the data repository; and
- Preparing this report.

Challenges encountered

By including the creation of an oral history archive as a goal from the beginning of the project, and therefore being able to explicitly ask the question in the consent form, the investigators in this project have been able to avoid ethical concerns about the archiving of the data. It has also meant that they have been able to create an archive where the bulk of the transcripts available have not been de-identified, and therefore create a richer and more valuable source of information for future researchers, as they will be able to put the responses provided into greater context. This has the flow on effect of reducing the risk of incorrect conclusions being drawn by future researchers relying on incomplete information.

Access levels

However, because the transcripts have not been de-identified, they do then need greater consideration of what access level they will be kept at in the ADA, to ensure that they are available for researchers, but that the participants are still protected. The transcripts cannot be made open-source as they do contain this sensitive information.

For the transcripts that have been de-identified (as far as possible – see discussion below), consideration then needs to be made as to whether a non-de-identified version of the transcripts should also be included without public access, and if it should be made available at a point in the distant future. The interviews in this archive were conducted between six to ten years ago,

meaning that seeking additional consent would be a challenging process, and, indeed, may not be possible for all interview subjects.

De-identification

For the ten interviews where permission was not given for the subject to be identified, a process of de-identification needed to be undertaken, to ensure that, as much as possible, the identity of the interview subject would not be easily discerned. While this consideration was absolutely paramount during the process, I was also endeavouring to ensure the transcript would remain a useful document for future researchers.

Furthermore, while the best possible effort was made to remove any identifying information from interview transcripts, the nature of an oral history interviews and the information provided in them means that, with sufficient contextual information, there always remains a possibility that someone could guess at who the interview subject is. For example, if a subject noted they worked at a hospital that specialised in a certain area, even with the name of the hospital de-identified, it could potentially be identified through this information. While this is a serious ethical consideration, in this context it is also one that the interview subjects were warned of, with the consent form noting that 'As the sample size for interviews is small, and given the nature of the interviews, it may be impossible to guarantee protection of my identity, even If I choose not to be identified, but every effort will be made to do so'. As such, there is no ethical bar against including these transcripts in the archive.

Below are the key de-identification protocols that I followed, including discussion of some challenging questions that arose as part of this process. As a standard point, all replacements in the transcript text were surrounded by square brackets.

1. Name protocol – interview subjects

For each interview subject, I replaced their first name with another beginning with the same letter. Some of the interview subjects already had pseudonyms created that had been included in summaries made publicly available – in those cases, I used the same pseudonym. For the remaining subjects, I used lists of popular names to select a name that evoked a similar age, ethnicity, and cultural association.

For their surnames, included in the header and introduction, I simply replaced these with the word 'Surname'.

Where the layout of the transcript included initials to indicate when the subject was speaking, I replaced these initials with the pseudonym chosen.

Case Study 1 – One interview subject briefly mentions a nickname given to them by the bus driver, prompting a question of if this should be changed. If a nickname were generally used, I would change it, likely with a similar substitution to those above. However, in this instance, as it was a name used only by one person, and totally unrelated to their own name, I have left it, as they are unlikely to be identifiable from this.

2. Name protocol – others

In the course of interviews, other names would be mentioned – those of teachers, students, family members, and colleagues, as well as more famous figures.

For the people that the interview subject was personally associated with, I followed the same protocol for changing first names. I replaced surnames with the first letter of the surname, or the first letter and an additional letter if there were two surnames beginning with the same letter. This was done for two reasons – firstly, because the people mentioned had not given their consent to be included; and because these names could identify the interview subject through association (e.g., the name of a school principal could identify which school an interview subject had worked at).

For people who were mentioned because they were generally well know – politicians, educational theorists, and others, I retained the name, as that could not identify the subject. Similarly, while school principals were de-identified, the head of a large organisation, if the organisation name was retained, was not. The exceptions made were for celebrities known for living in a certain area, or if they were mentioned because they had attended a particular school, as that could then identify the interview subject.

If the same name came up in different interview transcripts, I used the same pseudonym.

Case Study 2 – in one instance, a participant had identified themselves as belonging to a particular political party, the name of which I had substituted with '[political party]'. The interview subject then made reference to a leading figure in that political party at the time, leading to the question of if the name of the public figure should also be changed, as it would indicate the political party involved. However, following the protocols outlined above, the politician name would be simple to deduce, particularly as the conversation indicated that the party was in power at the time described. As such, I kept the politician name, and as a flow on from that, the name of the political party the interview subject was a member of. I did de-identify the names of the committee they worked on, though it is likely discernible from context.

Case Study 3 – One anonymous interview subject mentioned the names of other people interviewed, who agreed to being identified. They also suggested names of people to interview. From context, it appears that these are not necessarily people that they knew well or worked with, but rather people who were important in the development of educational and curriculum policy. As such, I have retained the names as I do not believe they indicate the identity of the interview subject, and could be useful information for future researchers in terms of people to consider.

3. Name protocol – locations

Location names associated with the interview subject, such as where they grew up, lived, or worked, were changed to protect the identity of the interview subject. Particularly for those subjects who lived or worked in small towns, giving this information could easily identify them.

When a subject spoke generally about a place that they were not associated with, the name of this place was retained, unless giving it would enable their location to be identified through process of elimination. This included if talks about the suburb that someone that the subject was discussing came from, when it didn't indicate the subject's own location.

I retained the name of capital cities. References to suburbs were changed to be 'Capital City Suburb #' (e.g. [Melbourne Suburb 1]; references to towns and cities (other than state capitals) were changed to be 'State Town #' or 'State City #' (e.g. [Victorian Town 2], [New South Wales City 3]). Each were numbered in order of when they first appeared in the interview transcript. If I was unsure if a location was classified as a suburb, town, or city, I went by the definition provided in Google.

Similarly, street names were changed to indicate they were a street or road. The formatting of this depended on what would be the clearest and least awkward in the text.

Case Study 4 – In one instance, an interview subject was talking of how siblings living at institutions could be broken up, with brothers going to one named location, and sisters another. Given the specifics of the place names given, there was a question of if they should be changed. However, as no specific children could be identified from this, the interview subject did not work at either of these locations, and providing the names provides accuracy in terms of the scale of the separation, the names were retained.

Case Study 5 – There was a question of if particular districts that interview subjects worked in should be de-identified or not. It is an area that the subject was working in, however districts would also be quite broad areas. Initially I changed the transcript when they noted they were working in the Western district, but there were also several passing references to working out west. The general location the subject was working in was also important context, in terms of indicating the socio-economic status of the schools they were working in. As such, there was an ethical question in balancing context and anonymity. In this instance, I chose to retain that they were working in the west, as it was broad enough to allow the context to be prioritised.

Case Study 6 – There were instances where the name should be de-identified, but it had not been recorded in the transcript, as it had been unclear. In that case, I left the markings from the transcribers, rather than inserting a pseudonym.

4. Name protocol – Institutions and organisations

The companies, institutions, and organisations that interview subjects worked at could be used to identify them, particularly smaller ones. As such, unless there were exceptional circumstances (see case study), these were changed.

For schools, hospitals, and other institutions that were named after the suburb/town/city the school was located in, I followed the location protocol outlined above, but retained the school type (e.g., [Sydney Suburb 4] Grammar, [South Australian Town 5] High, [Victorian City 6] Boys, [Adelaide Suburb 7] Hospital). The exception to this was schools named after the capital city, in which case the generic 'school name' was used. I had some concern that leaving a school type could potentially help in identifying a school name. However, the risk seemed minor, and the clarity provided by including the school type (particularly when several different schools in the same area were named) outweighed the risk.

For churches or schools named after a saint, I substituted with the name of another saint, ensuring first that it was a saint's name recognised by the particular denomination.

When schools, hospitals, and other institutions had a more specific name, I replaced with them with a generic 'School Name', 'Hospital Name', 'Company Name', followed by a number if there was more than one in the transcript.

Institutions were changed that the interview subject was identified with, and I also changed the institutions of minors mentioned by name, to protect their identity.

However, I have retained university names, as the large size of universities means that a person could not be identified by their attendance at any particular one, and the university attended could prove beneficial to a future researcher looking at the different approaches taught at different universities.

Case Study 7 – One interview subject worked for the organisation or government department that provided psychological services and guidance counsellors for all the schools in the state. In this instance, because it appeared to be a large organisation, with no clear substitute for the organisation name, and would be apparent from context where the subject was working, this was retained. Similarly, the name of the head of the organisation was retained.

In a similar case, the interview subject gave their role, of being a district guidance officer. While that could potentially be identifiable, as not many people were in that role, it is still a generic title as the specific district was not named, and there was no clear substitution that could be made.

Case Study 8 – An interview subject working with alternative schools named all the different alternative schools in the city/region, and later identified the two different models, and which schools fit into which model. To de-identify only the school they worked at would make it identifiable, and so these needed to be changed. However I had concerns regarding the clarity of the transcript, and its usefulness for future researchers if all the names were changed, particularly if a future researcher was unfamiliar with alternative schools. In this instance, I de-identified the names of all the schools using the model that the school the subject worked at was using, but retained the names of the schools using the other model.

Case Study 9 – An interview subject mentions both a Melbourne school that they were potentially going to attend but didn't, and hospitals where they applied to work but did not get in. With these I considered if they could be identified by these and therefore needed to be de-identified. However, because they were located in a capital city, where there are a large number of alternative schools and hospitals that they could have attended or worked at, I retained the names.

5. Other de-identification considerations

The process of de-identification is one of constant evaluation. Even when I had established a set of protocols for myself, these were constantly being challenged with new or slightly different circumstances. While retaining anonymity of the interview subjects (as far as possible) was central to this decision making process, as can be seen from the examples given above, there was constant consideration regarding the need to de-identify something, and, when it was a grey area, balancing the value of the information against the possible risk of identification.

Not every instance of de-identification is included in this document (there are over eight hundred changes across the ten transcripts in the record I kept), but the protocols and examples provided should give a sense of the approach taken. Below are some case studies of issues that came up that did not fit neatly into the protocols as described.

Case Study 10 – One interview was a husband and wife who were interviewed together, and so I have accounted for them as a group interview, though the record keeping does not list them as such, as they each completed individual consent forms, rather than group interview consent forms. The ongoing impact of that was that one partner stated that they did not wish to be identified, while the other partner stated that they agreed to be identified. In this instance, I have treated the transcript as a whole as one to be de-identified, for giving any information about the partner who agreed to be identified risks identifying the other partner. While this strongly seems to be the most ethical practice, it does have the negative consequence that someone who chose to be identified no longer has the opportunity of being identified as a contributor to the study.

Case Study 11 – One interview subject gives the name of a trophy presented at the school. The transcriber was not able to completely provide the spelling of the trophy name, however this has

still been changed to be [Trophy Name], as it is possible to determine its name, and therefore link it to the school and town they were located at.

Process of uploading the files

The process of uploading the files to the data repository shells was fairly straightforward, especially with the assistance of the staff at the ADA. There were some issues, particularly regarding zipping and applying a password using a Mac, and therefore the ADA staff being able to access them, and also the use of special characters not working in the password, but the staff were supportive, and also updated their documentation in the wiki based on our experiences.

The process of completing the metadata in the data shell was also fairly straightforward, though that was in large part because I had an existing metadata shell to base my answers off. I think some of the terminology may still be geared primarily towards quantitative datasets, though cannot specify any particular questions this was an issue with. This particular metadata set also complicated the process by having quite a large number of related publications. These have been entered in, along with appropriate DOI, ISSN or ISBN numbers into the publications, where possible.

Other comments

At present, the oral history archive is being restricted to transcripts only, and not the audio files of the interviews. If it is determined that these should also be included, serious consideration will need to be given to if the audio files of the de-identified interviews could be included. The nature of the de-identification process was such that there were redactions on nearly every, if not every, page of the transcripts. While the redacted transcripts could be used as a guide, the process of de-identifying the audio files would be onerous and potentially make the files unusable. This then leads to another ethical question of providing the audio files for the non-redacted interviews if we do not provide the files for the redacted ones, though at a surface level consideration the benefits would be such that they would outweigh any concerns regarding this.

Appendix 2: Globally Mobile Lives and Suburban Government Schooling

Project led and summary prepared by Joanne Higginson

Project summary

The project *Globally Mobile Lives and Suburban Government Schooling* forms my PhD project, currently in process (2019). I am investigating how biographical and intergenerational narratives might give us insights into social processes of "globalisation on the ground", especially within suburban government schools, where, I contend, transnational student experience of education and life across both Australia and other national contexts has become a more common and differently experienced phenomenon than in the past. My research questions are:

- In what ways might the biographical and intergenerational perspectives of young people and families who have experience of work and education across different national contexts assist in understanding some of the localised but potentially generalisable impacts of globalisation, including for 'local' schooling?
    - Relatedly: What are the limits and affordances of qualitative longitudinal and cross-generational 'biographical' interviews for exploring social change and change in process?
- How do place-based and transnational family experiences and family narratives influence identity formation and imagined futures for young people whose journey to Australia has been shaped by contemporary mobilities and migrations?
- How do these experiences intersect with contemporary 'local' school experiences in Australia?
- How is locality constructed and experienced in globalising times and how might this be reflected and refracted in school contexts?

Materials archived

I had participant consent for 15 of 30 interviews conducted for the project to be deposited with 'an archive service'. The interviews are with students from three different government secondary schools and their parents. In one case a parent provided consent for archiving, but her son did not. The interview transcripts vary in content and length, with parent interviews generally being longer and more detailed than those conducted with students. I have included a very brief description at the beginning of each interview indicating some of the issues and themes covered.

My original participant consent forms included options for interview transcripts to be digitally archived, for potential access by other researchers in the future. SOCEY had not been developed at the time I was seeking ethics approval; however, Julie McLeod's ARC-funded *Making Futures* project (ARC FT 110100646 Future Fellowship; *Making Futures* n.d.), with which my PhD is aligned, had included options, prompting consideration for this project too. I have deposited blank copies of my participant information, plain language statements and consent forms in the SOCEY archive as these may be of interest to others considering a similar approach. Sample interview questions are also included.

In addition to providing historic information about unfolding processes of globalisation and changing experiences of migration and schooling, the interview transcripts provide insights into student and parent uses of social media for transnational communication and family life. There are also references to specific historic events – such as the Greek and Venezuelan economic crises. I consider the transcripts as having historic as well as sociological value, something that I discuss from an epistemological, as well as practical, point of view below.

Archiving process

I transcribed all of the interviews myself – they were between 40 minutes and one and a half hours long. As is common practice, I had engaged in some anonymisation during my transcription (e.g., Clark 2006, 5). The main work I conducted in preparing my materials for archiving was checking and 'cleaning' my transcripts, ensuring consistent and archivable formats and preparing brief descriptors, plus a meta description of the data.

In terms of time and resources, I spent approximately eleven days:

- Re-checking consents and seeking additional consents (which were not forthcoming);
- Checking, cleaning (anonymising) and formatting interview transcripts prior to deposit;
- Liaising with the Australian Data Archive;
- Writing a data description; and
- Researching and writing this report.

Challenges encountered

Consent to archive

An initial challenge was that SOCEY had not been developed at the time I was negotiating consent for my participants and their interviews. At the time of recruitment the idea of archiving could only be discussed in general terms. I referred to socially and historically significant resources, such as the UK's Mass Observation Archive in some of my recruitment presentations and discussions with students.

A minor challenge was retrospectively seeking consent for archiving. Basically, all participants remained with their original decisions to have their transcripts archived or not. I believe that those who provided initial and ongoing consent for this option were motivated by ideas of historicity and were reassured that their transcripts would only be accessed for legitimate research purposes only – i.e., they cannot be uncovered via a simple internet search. This is an important consideration within a contemporary public culture where the volume and accessibility of information about individuals, including young people, has reached unprecedented proportions (e.g., Broom, Cheshire & Emmison 2009, 1178). For instance, Victorian Certificate of Education (VCE) results can be found through internet searches, and school newsletters are often available via school websites. This is in addition to social media presences through apps such as Facebook and Instagram. What I have taken away from this is that permission to archive may be best established in the initial relationship building and consent stages of the project – rather than through retrospective emails sometime after fieldwork has finished.

Reflexive and intuitive, versus blanket, anonymisation

In working through the challenges of anonymising my transcripts, a resource that I found very useful is Andrew Clark's (2006) paper *Anonymising Research Data*, written for the UK's Economic and Social Research Council (ESRC National Centre for Research Methods, NCRM

Working Paper 7/06). Clark argues for a 'more reflexive, iterative approach to anonymisation and confidentiality that situates these, and other ethical concerns, in the context of social processes'. Clark, citing Singleton and Strait (1999), argues that complete anonymisation may be impossible (Clark 2006, 4). Basically, I have intuitively and reflectively worked through five 'layers' of anonymisation. These layers and some of the associated process, considerations and challenges are outlined below:

1. Participant pseudonyms

I assigned pseudonyms to all my participants, some of them chose their own. Andrew Clark outlines some of the inherent challenges of this:

> Both personal (first) names and surnames imply particular ethnic, religious, class and age-based connotations, which will inevitably be transferred to any pseudonyms. (2006, 6)

A particular challenge I faced was that the majority of my participants were born outside of Australia and are from Language Backgrounds Other Than English. The Australian and UK data protection acts (*Privacy Act 1988* [Australia], *Data Protection Act 2018* [UK]) consider racial and ethnic origin to be 'sensitive data' warranting particular protective attention (Clark 2006, 7). I generally chose names that reflected both my participant's cultural backgrounds *and* their current Australian context – for instance, I chose popular, but hopefully not stereotypical, Indian names from the region in which they were born, for an Indian father and daughter, who have retained their Indian names in their new Australian context. Conversely, my ethnic Chinese participants had generally adopted Anglicised names in Australia – I chose similar (generally older) English pseudonyms. One father and son pair – whose interviews are not included in the SOCEY deposit, independently chose the English pseudonyms – 'William' and 'Harry'. I worked with 'Harry', but changed 'William', due to the perhaps accidental cultural connotations to the British Royal Family!

2. Pseudonyms for related parties who have not been part of the consent process

The process of anonymising necessarily extended beyond interview participants. I also took care to anonymise – sometimes using pseudonyms, sometimes using more general terms – references to family members, such as siblings, who were not interviewees and who therefore had not completed consent forms. Where possible I used general terms such as 'my other daughter'. There were also references to teachers and other students, which I had not particularly attended to during the transcription phase, which I was conscious of 'cleaning' for SOCEY.

I retained names for contemporary public figures such as Pauline Hanson, Sam Dastyari and journalist Sami Shah.

3. Considerations of space and place

Within qualitative research, spatial and place-based considerations may be intrinsic to the research, rather than simply 'background data' (Clark 2006, 5). This was particularly so in the case of my project, which foregrounds transnational social connectivities. Rather than adopting 'blanket anonymization strategies' I considered a combination of participant anonymity *and* research and analytical context and situatedness. My considerations and practices are outlined briefly, below:

Assigning pseudonyms to school research sites

In common with most contemporary educational research, I have used pseudonyms for the three schools from which I recruited participant students and families. I had given an undertaking to do this in my Department of Education and Training ethics application and in my communications

with school principals. There are characteristics of the schools, however, which may make them identifiable – probably more so in my PhD than in the SOCEY archive.

Anonymising most local place names

Within the broader *Making Futures* project, specific places and their histories are considered integral to the project and, through negotiated consent, have not been anonymised. In my own project, it is a type of school and location (mid socio-economic schools in culturally diverse middle-ring suburbs) that are considered significant. In light of this, I have anonymised most contemporary local place names and used general identifiers such as 'the local shopping centre' and 'my old primary school'. I have retained some local place names, especially where they relate to narrated pasts and/or provide important contextual information (e.g., Victoria Market, Luna Park).

Retaining most international place names within a project that focuses on transnational social fields and practice

The relationship between qualitative research, anonymisation and context is both practical and epistemological (Clark 2006, 12). In a project that focuses on migration and transnational family life, aspects of place are a key feature of the research purpose and context. I have retained as much information about international and overseas locations and contexts as possible, including in one instance, retaining the name of an overseas located school due to its unique characteristics.

4. Anonymising organisational names

The main consideration here was anonymising company and parents' employers' names, which could lead to identification of participants in an era of online professional networks, such as LinkedIn. Conversely, I retained some of the names of companies where students work – such as Kumon, Woolworths and KFC – but removed suburb names (most of these interview transcripts were not deposited with SOCEY). These are larger employers that employ significant numbers of young people and speak to different hierarchies and types of youth employment experience.

5. Removing reference to very personal experience and events

This is self-evident! I met with or communicated with some interviewees on multiple occasions and some were very frank in their interviews. I removed references to some difficult events, such as personal illness.

Benefits provided

My description of the potential benefits of archiving aspects of my project is written with a view to some of the issues deliberated and discussed by Broom, Cheshire and Emmison (2009) in their paper *Qualitative Researchers' Understandings of Their Practice and the Implications for Data Archiving and Sharing.* They note that epistemological issues associated with qualitative and interpretative social and educational research have historically been a sticking point in relation to data archiving and potential reuse. They report, for instance, that

> some have argued that research data derived from interpretive approaches in the social sciences typically involve subjectivities and epistemologies that do not lend themselves to data archiving … the practice of qualitative research is generally seen as one of 'generating' rather than 'collecting' data, with data being co-produced by the researcher and research participants. (Broom et al. 2009, 1164)

This sense of co-production is very evident in my interviews and there are some where I am very present in terms of generational, cultural and life experience as an interviewer (laughing for instance with one of my participants about the 1980s British television show *The Kenny Everett*

Video Show, which was a feature of both of our school days and cultures, or using my knowledge of my son's supermarket job to ask questions of a participant doing the same job, for the same company, in a different suburb).

Epistemologically I think of my interview transcripts, as co-constructed 'texts', which reflexively narrate personal histories and views. In this sense I think that Van den Berg's comments are useful, he argues that 'the empirical … is undoubtedly connected to the theoretical, but it also has a momentum of its own' (Broom et al. 2009, 1165). My focus is on transnational social practice, which will undoubtedly continue to travel in the future.

The historicity and subjectivities of interpretative qualitative practice (Broom et al. 2009, 1164) are, I believe, part of what make it so interesting and are not necessarily a barrier to archiving and potential future access and use. I was guided in this instance by thinking of the marginalia and notes within the Prest 1941–43 Social Survey of Melbourne (Wilfred Prest Collection, 1973.0002, University of Melbourne Archives) where the spatial and historical location and subjectivities of both the women university student researchers and their often working class female research subjects have informed rich, contemporary studies (e.g., Warne, Swain, Grimshaw & Lack 2003).

There are also paradoxes and pitfalls within our current digital public culture, which now need to be thought through in project conceptualisations. In researching potential schools to approach for my study, I attended a school open day which included a historical display. The display included printouts from previous decades of the school magazine. Among these were a series of profile interviews, conducted in the 1960s by a student with fellow students who were recent migrants, at a time when they were the same age as my study participants. These interviews are, in a sense, in the public domain, in that they can be found via the school's website, which has digitised archives of its school magazine; this was something that the interviewees could never have foreseen in 1968. I believe that a select, responsible, proactive approach to digital archiving is something that contemporary researchers now need to consider, especially for research using public funds and resources.

I am guided most of all by senses of collective knowledge and historicity in researching social change.

Other comments

My experience is that consent to archive is best negotiated in the early stages of a project. As SOCEY had not yet been developed at the time I conducted my participant recruitment and interviews, I discussed the idea of archiving in general terms – referring to the Mass Observation Archive, for instance, in my recruitment presentations. This may have deterred some students from stepping forward for the project, however, it may have interested others. I believe that it is wise to include options for archiving, even when you don't know what those options might be. Consent to archive varied across schools – for instance, at one school only a minority of participants gave consent; at another, almost all participants did. I did not probe this difference at the time.

I believe that there are social and historical synergies between the projects currently included in SOCEY. To use a digital metaphor – I think of them as a well curated *Spotify* play list! Each is independently significant, but they also 'hang together', providing a broader and richer sense of the scope and synergies of social and qualitative research within this moment.

References

Broom, A., Cheshire, L. and Emmison, M. 2009. Qualitative Researchers' Understandings of Their Practice and the Implications for Data Archiving and Sharing. *Sociology* 43.6, 1163–1180. DOI: 10.1177/0038038509345704

Clark, A. 2006. *Anonymising Research Data*. Working Paper, Economic and Social Research Council, National Centre for Research Methods, Working Paper Series 7/06.

Warne E., Swain, S., Grimshaw, P. and Lack, J. 2003. Women in Conversation: A Wartime Social Survey in Melbourne, Australia, 1941–43. *Women's History Review* 12.4, 527–46.

Wilfred Prest Collection, 1973.0002, University of Melbourne Archives.

Appendix 3: Childhood Maltreatment and Public Inquiries

Project led by Katie Wright. Summary prepared by Sari Braithwaite and Katie Wright

Project summary

This project is an historical sociological study that examined the unfolding of the Royal Commission into Institutional Responses to Child Sexual Abuse (2013–2017) alongside past Australian inquiries (1970s–2000s) that either focused explicitly on child maltreatment or included this within their remit. The project also situates Australian inquiries within a wider international context.

The aim of the project was to explore the role of public inquiries in changing understandings of children's development, vulnerability and rights, and in turn how this has shaped social policy, educational responses, and public attitudes towards safeguarding children and promoting their wellbeing. A key focus was investigating how concepts of childhood and policy approaches are changing as a result of an increasing social imperatives for openness and disclosure about matters once considered taboo, including child sexual abuse. Overall, the project sought to advance conceptual policy insights on a major social issue and sociological insights on childhood and the forms and effects of late modernity.

This project was funded by the Australian Research Council, DE140100060, 2014–2019.

Materials archived

This archive is comprised of transcripts from key informant interviews undertaken for this project.

Most interviews were conducted in Australia, but the collection includes interviews that took place in England, Northern Ireland, Ireland, and the United States.

In Australia, participants were interviewed in New South Wales, Victoria, and Queensland.

Participants include survivor advocates, commissioners and staff members of inquiries, lawyers, and journalists.

At this stage, we are archiving interviews with 23 participants. We anticipate a smaller, second round deposit in 2020.

It was intended from the beginning of the project that an archive would be created for future researchers. The Chief Investigator, Katie Wright, included questions in the consent form for all interview participants regarding permission to deposit their interview in an archive.

Twenty-three participants indicated that their interviews could be accessed for future researchers. One participant wanted identifying information to be removed and with one person we need to clarify redactions/amendment and permissions.

There were four interviews which were not recorded, as this was the preference of those participants. Notes from these interviews have not been included in the archive.

Short summaries were created for each interview, with biographical information about the participant, and a summary of key discussion points, topics covered, and organisations mentioned. The summaries also have a list of keywords which can be entered into metadata.

Supplementary material, consisting of Participant Information Statements and blank Consent Forms, were compiled into a single document for archiving.

Work required to develop the archive

The interviews were conducted by Katie Wright. Transcripts were prepared from the interview audio by a professional transcriber throughout the project.

Participants had the option to review the transcript of their interview, with edits being made and reviewed as part of that process.

If a participant made amendments or edits to their transcript, these were accepted and deposited as the archival version.

Biographical information and summaries were also created for all interviews.

Sari Braithwaite was engaged to prepare the materials for archiving. This work has included:

- Checking consent forms to ensure they matched the consent levels recorded;
- Contacting participants regarding clarification on their consent;
- Contacting participants regarding the editing and returning of transcripts;
- Checking each transcript to ensure edits had been completed, correcting minor errors found, and attempting to fill in 'inaudible' parts of the transcript;
- Formatting all interviews with consistent design and information in titles and headers;
- Writing summaries of each of the interviews;
- Adding biographical information to the summaries – and, where possible, liaising with participants to ensure the information was accurate;
- Identifying which of the transcripts were not be archived;
- Liaising with the Australian Data Archive (ADA);
- Entering metadata and data description into the data repository shell created;
- Formatting and uploading the transcripts to the data repository; and
- Drafting this report.

Challenges encountered

The goal from the beginning of the project was that interviews would be archived at the completion of the research. The investigator was thus able to avoid ethical concerns about the archiving of the data with information about this provided to the participant at the time of the interview and consent provided either at that the time of interview or after participants had reviewed their transcript.

However, we still had a number of challenges that we needed to navigate.

What to archive

This is the biggest question we had in preparing for this project. We started with a huge list of materials which could potentially be archived: all supporting ethics and administrative documents, research profiles, questions for interview participants, individual correspondence with participants, individual consent forms, original audio recordings, original transcripts and edited transcripts.

We undertook a process of interrogating what could be deposited, asked ourselves why it would be deposited (what value does it have?) and what were the ethical implications of what we were doing. A major consideration was consistency across the dataset – we wanted to deposit a consistent set of data which made the archive clear to use and easy to understand.

De-identification

For one interview, permission was not given for the participant to be identified, and a process of de-identification will need to be undertaken. We did not have the capacity to navigate this challenge before this deadline; this will be something that we will address before our second deposit.

It was very helpful to review Emily Fitzgerald's report [Appendix 1] on the protocols she developed for de-identification, which included stripping information about name, location, institution and organisation. In the case of our interview, the participant has indicated that we can identify them as a staff member of a particular institution – which means at this stage, we will need to work with the transcript to de-identify their name and role at the organisation. We may also send the de-identified transcript to the participant for their approval before final deposit.

Incorrect completion of consent forms

In the design of the consent form, there may have been too many questions, which resulted in a small number of participants either skipping sections of the consent form, or incorrectly filling it out (e.g., ticking all boxes, when they were being asked to make one selection). As a result, it was difficult in some cases to ascertain what levels of consent had been granted. We were able to contact participants in some cases and clarify; in other cases, we knew we met the conditions to archive for research purposes and could still proceed with the deposit.

Zipping files with a password on a Mac computer

The ADA instructions were very helpful, and the SOCEY team were great in responding to any questions. Zipping files with a password on a mac computer proved difficult. A way of doing this within the terminal was possible; however, the process is rather cumbersome, taking around two minutes per zip. If we had hundreds of files that needed to be deposited, we would have needed to keep searching for a different way of doing this.

Case Studies

<u>Case Study 1: De-identification</u>

In one case, as noted above, the participant gave permission for their interview to be archived but did not want to be identified. However, the process of de-identification in this case is not straightforward, given the participant's employer and their role within that organisation. It is possible that the amount of information required to de-identify this transcript will be so extensive,

due to the amount of contextual information that will need to be removed, that it may render the interview unsuitable for archiving.

Case Study 2: Off-the-record comments

In one interview, the participant noted in a number of places that what they were saying was not to be part of the official record of the interview. However, upon reviewing the transcript, it is difficult to ascertain whether it is only the sentence that should be removed, the name of a person, or the entire anecdote. In the case of this interview, we need to discuss each instance with the participant to clarify what they approve being in the final transcript. An issue for consideration here is the time imposition on the participant. All participants were given the opportunity to make amendments to their transcript, which requires time. For this interview, clarifying what should be included requires additional time.

Other comments

Question regarding audio archiving

Emily Fitzgerald indicated a challenge that we too encountered in regards to depositing audio files. From her report [Appendix 1]:

> At present, the oral history archive is being restricted to transcripts only, and not the audio files of the interviews. If it is determined that these should also be included, serious consideration will need to be given to if the audio files of the de-identified interviews could be included. The nature of the de-identification process was such that there were redactions on nearly every, if not every, page of the transcripts. While the redacted transcripts could be used as a guide, the process of de-identifying the audio files would be onerous and potentially make the files unusable. This then leads to another ethical question of providing the audio files for the non-redacted interviews if we do not provide the files for the redacted ones, though at a surface level consideration the benefits would be such that they would outweigh any concerns regarding this.

Question regarding biographical information

It was only at the end of the project that we realised that biographical information on interview participants would be helpful. In the future, as part of the consent process at the time of the interview certain biographical details should be discussed and confirmed with the participant. We could provide the participant with biographical information that we gleaned from our research – and confirm those details with them, or allow them to make changes.

Appendix 4: The Curriculum Policies Project

Project led by Lyn Yates. Summary prepared by Henry Reese

Project summary

School Knowledge, Working Knowledge and the Knowing Subject: A Review of State Curriculum Policies 1975–2005 was an ARC Discovery Project that ran between 2007 and 2008. The University of Melbourne funded a further year of research in 2009. The Chief Investigators on this ARC grant were Lyn Yates and Cherry Collins of the Melbourne Graduate School of Education, with additional research assistance provided by Kate O'Connor, Katie Wright and Brenda Holt.

Responding to a noted dearth of systematic scholarship about the development of state curriculum policies, the Curriculum Policies Project aimed to produce a foundation picture of developments in curriculum policies across the nation over a 30-year period. The project provided a wide overview of the last generation of state curricula, moving past previous projects that were limited in scope to individual government reports, Commonwealth developments, subject areas or political contexts. The overarching focus of the project was on charting continuities and changes in state curriculum policies, especially regarding changing approaches to knowledge, to students, and to the marking out of academic and vocational agendas. The focus was broadly on secondary schooling, and aimed at building up snapshots of curriculum changes at ten-year intervals.

This Curriculum Policies Project comprised two broad research tasks. The first was to compile an overview of resources that other scholars and students could access, bringing together the relevant chronologies, key documents and political background to the changes in state curriculum policies over the thirty years in question. This information can be accessed via the project website.[8] Changes *over time* and *between states* are mapped alongside each other to give a synthetic picture of changing Australian curriculum policies since 1975.

The second major research task – and the subject of this report – was to supplement this research on formal policy changes with a series of oral history interviews with key figures who had a say in the development of these state curriculum policies since 1975, either through shaping policy or in researching education. Thirty-four public servants and education department officials, curriculum academics and scholars were interviewed by Lyn Yates and Cherry Collins over 2007 and 2008. Interviewees were asked to give their personal reflections on the broad changes in curriculum policy over the years in question, and were invited to shed light on the reasoning and institutional factors that lay behind various policy decisions. The interviews were broad-ranging, informal and largely open-ended; research participants were asked to give a general assessment of their own involvement in curriculum over the 30 years in question, and to highlight any landmarks that were significant to them. They were also invited to address the broader themes of the research study, namely changing attitudes to knowledge, to students and to academic/vocational agendas, and to similarities and differences between different the approaches taken in different states.

[8] *Curriculum Policies* project, n.d., https://education.unimelb.edu.au/research/projects/curriculum_policies_project

Taken together, the interviewees' personal impressions and frank recollections of successes and failures, regrets and predictions, supplemented the hard policy data obtained as part of the project. This sheds further light on some of the broad overarching principles regarding knowledge, students, and academic and vocational agendas that came into play at critical moments in the development of state curriculum policies. This qualitative data provides rich insight into the institutional contexts, the idiosyncrasies and local factors that obtained in individual state contexts, and the personal relationships that were all critical to the broader policy changes charted over the thirty years in question. The rich and informal interview transcripts also contain valuable insights into the social history of childhood and education in postwar Australia, the changing role and status of the professional 'expert' in Australian public service, and the political context of education policy, which has attracted significant media attention at landmark periods in the study.

Materials archived

The standard interview consent form required participants to respond to two questions. The first concerned consent and confidentiality:

> I agree that comments made in my interview may be quoted and that I may be identified as the source of these, except where I indicate orally during the interview or in subsequent comments on the transcript that I do not wish to be so identified.
>
> OR
>
> I agree that my interview may be drawn on in the overall research project, but it should not be quoted or used in ways that identify me as the source unless I give specific subsequent permission to do so.

The second concerned retention or destruction of materials at the end of the research project:

> At the completion of the project (plus five years) I wish my interview tapes and transcripts to be destroyed.
>
> OR
>
> At the completion of the project, I consent to the placing of my interview tapes and transcripts in an archive that may be used by future researchers.

Of the 34 state curriculum experts interviewed for the Curriculum Policies study, 19 gave consent for their interview transcripts to be archived. In total, 17 interview transcript files were deposited with the Australian Data Archive (ADA).[9] Seven participants did not give consent to be archived, and a further eight transcripts did not have a consent form on file. It was decided that these could not be archived as consent to archive could not be positively determined. This is unfortunate, as the consent forms for all six South Australian interviewees are lost, meaning that this state is not represented at all in an archive which aims for national scope.

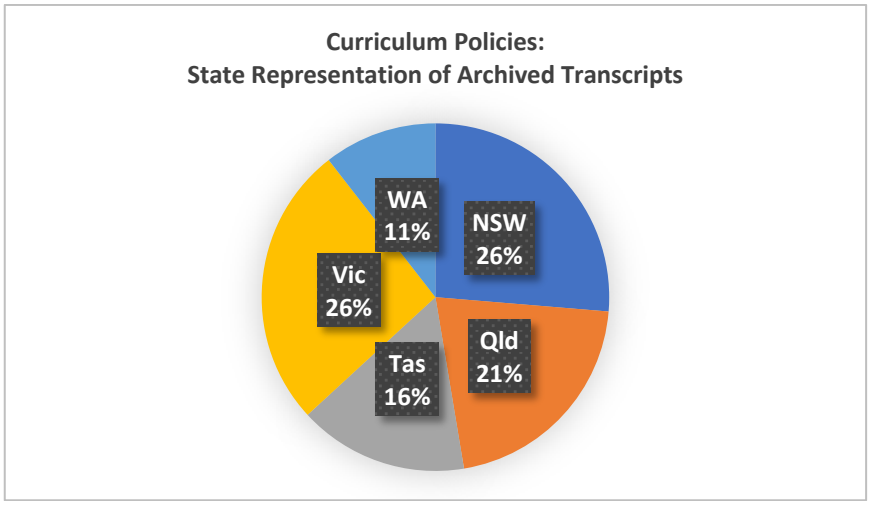
In addition, the Chief Investigators' interview summaries, containing brief summaries of the interviews on a state-by-state basis, were also archived. An 'archival overview' document also provides a summary list of the name and roles of each archived interviewee. Copies of the consent forms for each interviewee were also included in the archive.

Supplementary materials that were also archived include the blank consent form and Plain Language Statement for the study, as well as a spreadsheet that provides an overview of the consent form data at a glance.

[9] Two of the archived transcripts are of interviews with two participants at once. As such, the 17 transcripts represent all 19 participants who consented to their interviews being archived.

By state, the archived interviews represent the following: five interviewees for New South Wales, four for Queensland, three for Tasmania, five for Victoria, and two for Western Australia (out of a total of 19 transcripts).

Two interviewees from New South Wales, two from Tasmania and three from Western Australia did not give their consent for their transcripts to appear in the archive. Accordingly these transcripts were not archived. In addition, consent forms were unavailable for six interviewees from South Australia and two from Western Australia. As a result, South Australia is unrepresented, and Western Australia is underrepresented, in the archived data. The following graph shows the state representation in the archived data:



As might be expected, the two most populous states, New South Wales and Victoria, are the most heavily represented in the data. The Northern Territory and Australian Capital Territory were not included in the study.

Work required to develop the archive

Firstly, I collated all available consent forms and tabulated them. As outlined above, this table appears in the dataset as a guide to the data. I checked and standardised all name spellings and interview dates, sometimes having to confirm online the correct spelling of an interviewee's name where different spellings appeared through the materials I was working with. From this I determined which transcripts were to be archived, guided at all times by the consent of participants (see below).

The next task required reading over and 'cleaning' the 17 interview transcripts to be archived. This encompassed several steps. Firstly, I standardised the formatting of each transcript to ensure uniform font size, title, naming conventions and spacing in each document. Next, I standardised the formatting (turning all double spaces after full stops into single spaces, reducing redundant ellipses and multiple exclamation marks, for instance) ensuring that the documents presented as a neat, cohesive whole. I ran an extensive spelling and grammar check on all transcripts to ensure that grammatical conventions were followed consistently, mindful that the transcripts will be deposited in permanent form online. I was careful not to intervene in an editorial manner in the text of the interviews themselves, which were informal and conversational, and transcribed by a third-party transcription service. As such, some words are misspelled or misheard, and included incidentals such as phones ringing, unfinished sentences, abbreviations and so forth. The live, acoustic character of the interview context is preserved in the transcripts, but this is not

inconsistent with the necessity for the documents to appear in a consistent and professional manner in the archive.

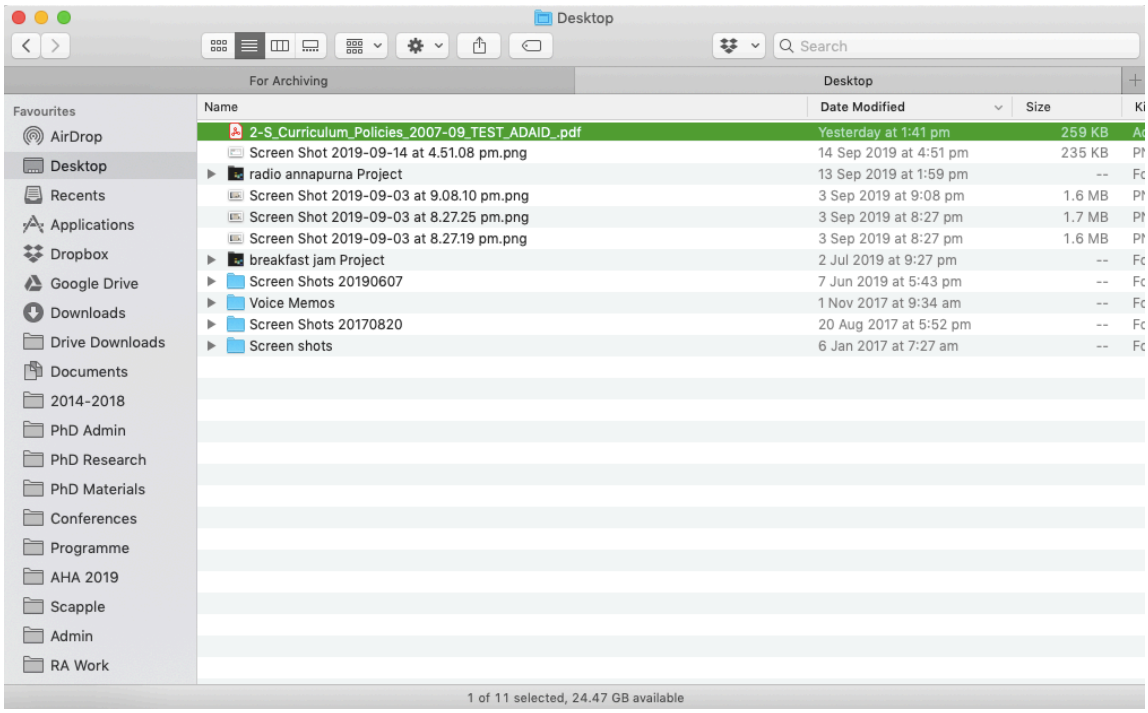
Next I prepared each of the transcripts for archiving, in accordance with the principles set out clearly in the ADA wiki (ADA 2019). This involved running a Norton Antivirus scan on each file to determine that no viruses or malware were present in the files, then saving each transcript as a writable PDF file. New names were created for each file in accordance with the ADA's naming protocols.

The next step was to zip and encrypt each transcript file for upload to the ADA Dataverse. The ADA wiki recommends the program '7-zip' for creating encrypted, password-protected .zip files. This program is not available on Mac, but thankfully Mac computers have the in-built ability to create encrypted .zip files. The process is straightforward, although it requires entering commands into Mac's Terminal utility and can be bewildering for first-time users or users without strong computer skills. I have outlined the process that I used in full here, in case it helps future ADA users unfamiliar with the process.

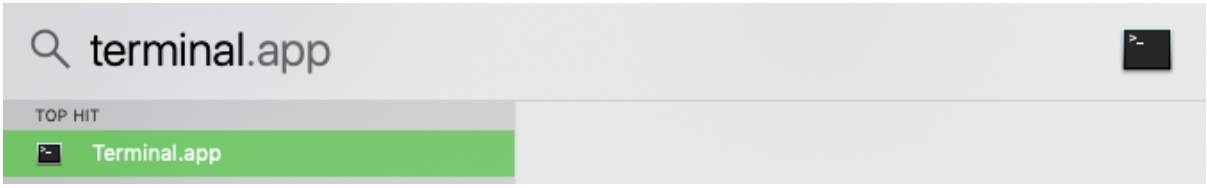
Zipping and encrypting files on Mac

Firstly, I copied every file that I wanted to encrypt onto my computer's Desktop. This makes it easier to enter the encryption command into Terminal, as Terminal must locate each file in order to perform this process.

For this example, I will be encrypting a PDF entitled '2-S_Curriculum_Policies_2007-09_TEST_ADAID_.pdf'. I have moved a copy of this PDF to my Desktop in preparation. Here is how the file appears in Finder:



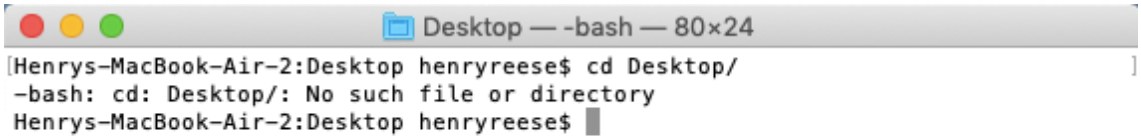
We then need to open Terminal and start the encryption process. You can find Terminal by pressing Command-Spacebar and typing Terminal into the spotlight search bar that appears.



Once you have opened Terminal, enter the following command:

cd Desktop/

Then press [enter]



This identifies the Desktop as the folder we are working in.

We now need to instruct Terminal to perform encryption on the relevant file(s). For this stage, it helps to copy your filename to the clipboard, ready to paste straight into Terminal.

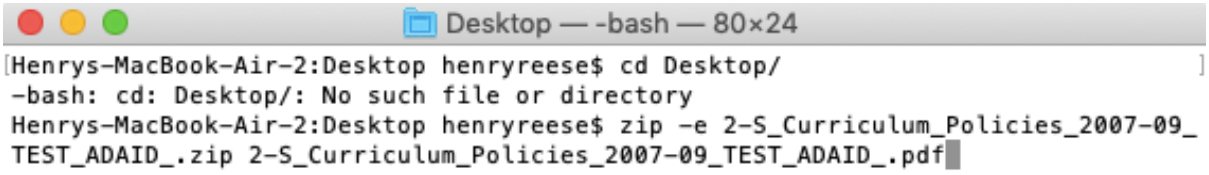
Enter the following command:

zip -e [filename].zip [filename].pdf

(note the spaces in the above command)

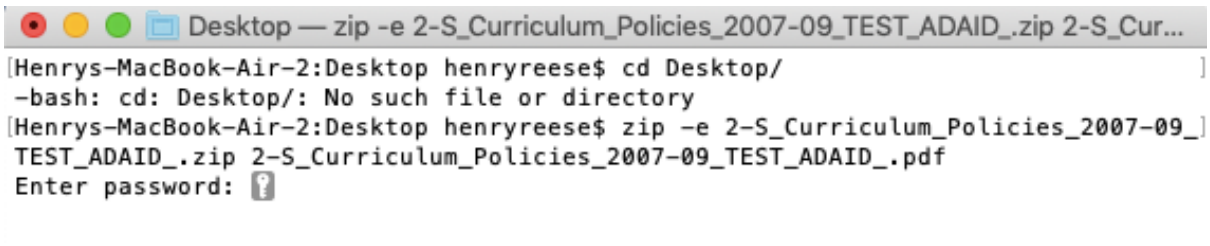
This tells Terminal to create an encrypted .zip file from the identified .pdf file. In my case, it looks like this:

Zip -e 2-S_Curriculum_Policies_2007-09_TEST_ADAID_.zip 2-S_Curriculum_Policies_2007-09_TEST_ADAID_.pdf



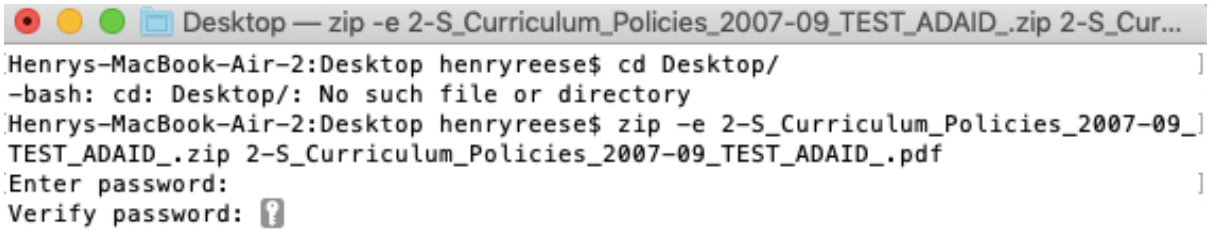
If you are encrypting a different kind of file, add the relevant extension to the second filename instead of .pdf (e.g., .xls, .csv, .rtf)

Then press [enter]. Terminal will then prompt you to assign a password to the file.

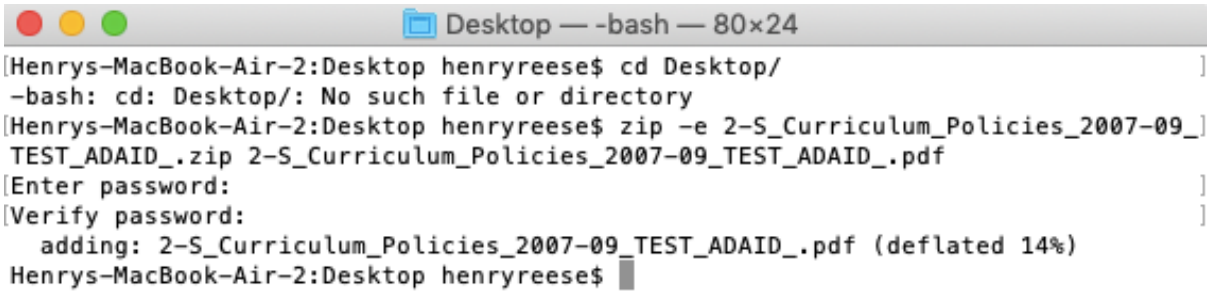


Enter your password and press [enter]. Note that you will not see your password as you type, but Terminal is monitoring your keystrokes. Be careful not to make any mistakes in entering the password, as you could end up with an unusable .zip file.

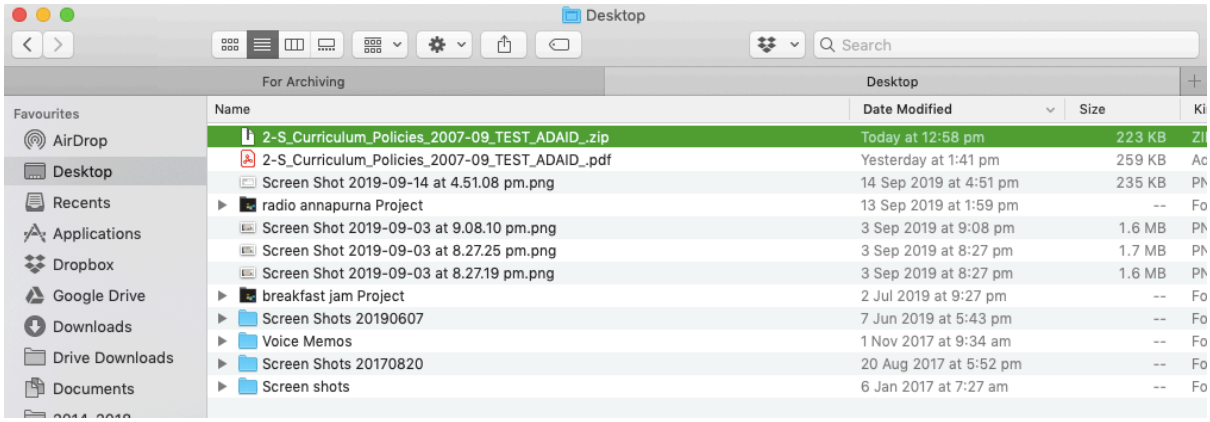
After entering your password, you will then be prompted to verify it. Enter your password again and press [enter]:



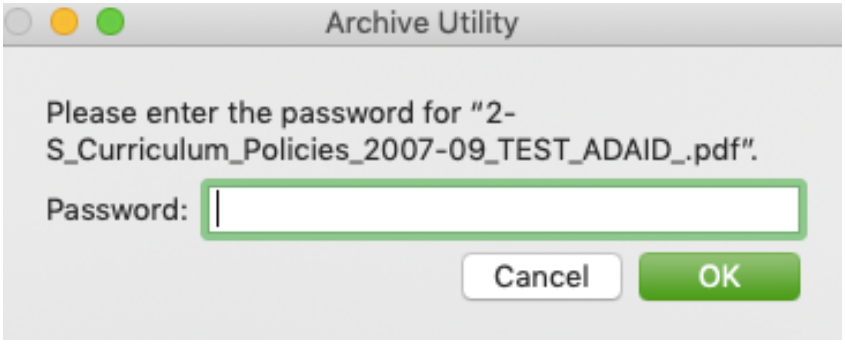
Terminal then performs the encryption:



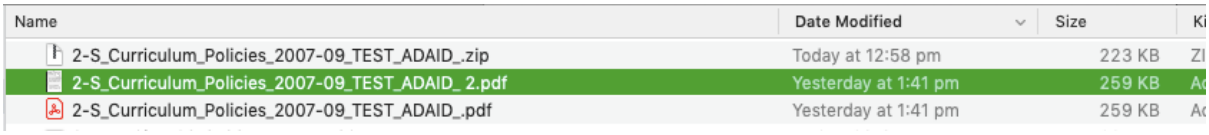
If you check your Desktop in Finder, you will find that a .zip file has been created with the filename that you specified:



You can test the encryption by double-clicking on this file. You will be prompted to enter the password that you just assigned:



Entering this password will extract the original PDF from the encrypted .zip file. It will then appear on your Desktop too:



The encrypted .zip file is now ready to upload to the ADA Dataverse, according to the process set out on the ADA wiki.

The final step in the archiving process was to upload all relevant documents to the ADA Dataverse, and standardise the metadata for each to ensure maximum searchability in future. Here, again, I was guided by the detailed guidelines set out in the ADA wiki.

Challenges encountered

Consent and confidentiality

The main challenges regarded interpretation of the consent forms. The consent form text (outlined above) is vague and badly formatted, and participants' wishes regarding future data usage could not be positively identified in some cases.[10] In these situations it was crucial to err on the side of caution, and to not proceed with archiving unless consent was clearly documented on file.

On several occasions, the consent forms were inconclusive regarding participants' consent to archive in the first place. Interpretation of the interviewees' consent form responses, over ten years after the conclusion of the study, often proved difficult. Five interviewees merely left Section 2 of the consent form (regarding consent to archive) blank, rather than positively indicating whether they wanted their transcripts to be destroyed or archived. I took these blank forms to mean that the interviewees did *not* wish to be archived, as did Lyn Yates in her original notes. But it is unclear whether, if they did *not* want their responses to be archived, this means that they *did* want their responses to be destroyed. In any event, the subjects did not positively give their consent to be archived. As a result, they were not archived.

Some questions remained. For instance, is it acceptable to include the names of participants who did not wish to be archived in the archived spreadsheet list of interviewees? Or are we to understand that these interview participants did not want their names associated with this study at all? It is impossible to determine this matter positively without referring back to the participants themselves. Mindful of how difficult it is to work out consent in retrospect, Kate O'Connor and I decided to include the names of those who did not give consent to be archived in the table of

[10] The consent form appears as a large, undifferentiated block of text. The different sections could have been demarcated much more clearly.

consent forms, although the matter remains unclear. Problems like this could be avoided by having clearer options on the consent form; rather than 'I would like my tapes destroyed' or 'I would like my transcript placed in an archive after five years has elapsed', an initial section could ask, in simple yes/no fashion, whether a respondent would like their transcript to be archived. Following this response, further options, such as destruction of data, could then be covered in more detail. Further, as well as asking whether clients would like to be *quoted* in the study, it is also imperative to ask whether they would like to be *identified* as involved at all; this is especially important given that the interview participants are public figures, and many of the interview transcripts refer to the work of other interview participants, including some who did not give their consent to be archived. It is accordingly impossible for these figures to be completely anonymous in this study. A redesign of the consent form would make this thorny process of reconstruction of participant consent much clearer. Future qualitative studies should bear this matter in mind in the construction of Plain Language Statements and consent forms.

A second, related challenge regarded interviewees' consent to be quoted in the study. If a participant gives their consent to be archived, but does not give consent to be 'quoted', does this refer only to publications produced as a direct result of the Curriculum Policies ARC project, or are we to take this as meaning they want their transcript to be anonymised for future archival access? Again, this problem would be obviated by a clearer consent form, with extra questions that cover more specific future situations. For instance, a participant could specifically indicate whether or not they wanted to be quoted (1) in the study, *and* (2) in the archive. Again, this is a valuable lesson for future researchers working with qualitative data dealing with public figures in a similar manner: it is important to ensure that every foreseeable future data outcome is included in a consent form, baked in from the start, as reconstruction over a decade later is difficult at best.

Of the 19 interviewees who agreed to be archived, seven did not give their consent to be quoted in the study. Upon consultation with Lyn Yates and Kate O'Connor, we decided that this response to the consent questionnaire did not preclude the participants' identification in the archive, i.e., that we can take this response as meaning they did not want to be quoted in any immediate publications that arose from the Curriculum Policies Project, but that it is acceptable to archive their transcripts without anonymisation. Consequently, no interview transcripts were fully anonymised. I closely checked each of the relevant transcripts for evidence of any understanding to the contrary and found no evidence of any understanding that their transcript would be made entirely anonymous. The wider implications of this matter are discussed below.

Qualitative data involving public figures

For this study, it is important to think deeply about *why* the participants may or may not have wanted to be quoted. This is a critical challenge that arose on several occasions, especially given that this study dealt with adult, professional public figures speaking about recent, often politically controversial topics in their portfolios. At the time of the study, the culture wars of the early 2000s were still raging over the status of education, particularly over the proposed Australian Curriculum, which was being developed at the time. Conservative educationalists such as Kevin Donnelly were publicly criticising the work of public servants and curriculum professionals in the mainstream media, many potentially defamatory accusations were publicly aired, and acrimony was rife in the sector. Many interviewees made mention of the fact that curriculum policy is a potentially controversial matter, and showed an understanding that this was a constitutive context in which they worked. This is another reason why the Curriculum Policies Project is itself so valuable in reconstructing the wider political and social contexts under which education policy developed. In this context, it is understandable that several of the key movers in the previous generation of curriculum policy did not want their names – and their candid opinions about

colleagues and other public figures – associated with their comments in the heat of this historical context.

At the same time, full anonymisation would simply not make sense in a qualitative study of this kind. The interviewees' very publicity is what is important here; they were the purveyors of first-hand, expert knowledge of the developments in state curriculum policies since 1975, and were interviewed on this basis. A balance therefore had to be struck between preserving the interviewees' wishes on the one hand, and ensuring the interpretive richness and consistency of the archive on the other.

In addition to the matter of quoting and anonymisation discussed above, I found it necessary to closely read through each transcript in order to identify if there were any specific instances where an interviewee indicated that they wanted to give an 'off the record,' without prejudice or confidential comment. I found five such instances, which are illuminating in regard to the breadth of potential confidentiality issues.

On one occasion, an interviewee referred to a private 'health problem' regarding the education minister in their state: 'Basically she had a bit of a breakdown, and she was definitely not on the ball …' The discussion of the health of another public figure is a highly sensitive matter, but given that this was a matter of public record, and that the interviewee expressed no personal reservations in stating this on tape, Kate O'Connor and I decided that this would remain in the interview transcript. This matter demonstrates, perhaps more clearly than any other example, the sensitivity of the archive's contents.

A second instance finds an interviewee expressing concerns that their comments on a certain policy change 'probably could be used for blackmail purposes, don't quote me'. Here, in line with the interviewee's clearly expressed wishes, I redacted the section that they referred to. Likewise, a third interviewee asked that their recollections of an overseas trip be made 'off the record'; this was a clear occasion where redaction was necessary to conform with the interviewee's wishes. On the final two other identified occasions, interviewees expressed strong opinions about the capabilities of their colleagues; one questions what a consultant is 'intellectually capable of' while another refers to a leadership figure as 'totally self-important'. Again, on each occasion they requested that their personal comments be 'off the record.' As such, these were redacted.

Given that I am not an expert in curriculum and had no personal relationship with any of the interviewees, it was essential to be guided by the transcripts: where an interviewee indicated they wished to be confidential, I preserved their wishes. Given the sensitivity of the materials in question, it was necessary to take the time to read each transcript closely to ensure that nothing was in contradiction to the express wishes of the interviewee.

Other comments

Sensitive information classification in the ADA Dataverse

Another observation pertinent to the non-anonymised data here is the question of the classification of the data according to confidentiality. The ADA wiki provides that the prefix 2-S (rather than just 2) can be used for datasets that contain sensitive information, and provides helpful definitions of personal and sensitive information as covered in legislation. Given that the data in this project clearly falls under these definitions, I used this appellation in the naming of the data.

This appellation was also used for all supplementary files that mention interviewees' names, including the following:

- 2-S_Curriculum_Policies_2007-09_Archival_Overview_ADAID_.zip
- 2-S_Curriculum_Policies_2007-09_Interview_Summary_Index_NSW_ADAID_.zip
- 2-S_Curriculum_Policies_2007-09_Interview_Summary_Index_Qld_ADAID_.zip
- 2-S_Curriculum_Policies_2007-09_Interview_Summary_Index_Tas_ADAID_.zip
- 2-S_Curriculum_Policies_2007-09_Interview_Summary_Index_Vic_ADAID_.zip
- 2-S_Curriculum_Policies_2007-09_Interview_Summary_Index_WA_ADAID_.zip
- 2-S_Curriculum_Policies_2007-09_Consent_Forms_ADAID_.zip

I feel that this is necessary to acknowledge that there is potentially sensitive data in every single transcript, and to therefore ensure that access is never completely open, but that it is managed appropriately by the ADA. Again, as with the other confidentiality matters outlined above, I feel that it is necessary to err on the side of caution here.

Appendix 5: Making Futures

Project led by Julie McLeod. Summary prepared by Monika Popovski

Project Summary

Making Futures: Generational Change, Youth Values and Education is a qualitative study of young people's journeys through the senior years of secondary schooling and into the world beyond. It explores how young people, living in contrasting regions in Australia, navigate their educational, social and familial worlds and imagine and work towards their futures. The study aims to yield insights into the contemporary experience of senior schooling and the socially diverse ways in which young people make their lives, and engage in ethical thinking about social inequality, citizenship, knowledge that matters to them, and social and personal values in interaction with schooling, families and location.

In particular, the project seeks to gain insight into their perceptions of gender relations and perspectives on diverse forms of social issues and differences. Themes examined include social differences, gender relations, immigration, place and identity, religion and everyday ethics. Parents are also interviewed separately about these matters and their own educational memories, values and future hopes when they were at school, as well as hopes for their child's future. As such it takes a cross-generational and comparative approach to processes of youth identity and educational change: comparing young people's perspectives today with their parents' recollections of their own educational experiences and attitudes to similar issues. The project's design provides immediate generational comparisons between parents' recollections of their education and growing up with children's contemporary experiences and views. Additionally, the project compares this current research with findings from and analyses of earlier studies of young people and schooling undertaken during the 1950s to 2000s.

The project research is embedded within communities of Collingwood, Warrnambool and Manor Lakes, in Victoria, and considers historical and social factors that impact on schooling and youth identity within them. It holds that the past lives of a locale, even the distant past, impacts on both its present and future. This draws on approaches to place-based ethnography and historical sociology, including social reputation, patterns of migration, employment, class status and educational provision, as well as notions of temporality and place.

Materials archived

A total 206 files have been archived. Seventeen of these files are Research Materials (filename beginning with 1); 54 are interview transcripts and one is interview metadata (filename beginning with 2); and 134 files are signed consent forms, original non-anonymised interview transcripts and an interview metadata file containing identifiers (filename beginning with 3). Files beginning with 3 will be archived only as part of the Submission Information Pack (SIP) and will not be available in published dataset. Please see below for more details.

Interview transcripts (of a total of 70, 54 are available in published dataset)

A total of 28 participants; eight parents/guardians and 20 students were interviewed. Each parent/guardian was interviewed once and students were interviewed one to four times, one

interview per year from 2015 to 2018. One parent and six students did not consent to archiving of interview transcripts so that a total of 17 interview transcripts have only been archived as part of Submission Information Pack (SIP) but not available in published dataset. Such files are identified by a 3 at the beginning of their file name.

- *Collingwood* – 20 interviews, 8 participants; 3 parents and 5 students
- *Warrnambool* – 27 interviews, 10 participants; 3 parents and 7 students
    - 1 student did not consent to archiving of interview transcripts (3 in total)
- *Manor Lakes* – 23 interviews, 11 participants; 3 parents/guardians and 8 students
    - 1 parent and 5 students did not consent to archiving of interview transcripts (13 in total)
- One spreadsheet file detailing participant data including biographical information

Research Materials archived begin with Dataverse number 1 (17 in total)

- Blank copies of interview questions
    - Students
    - Parents
- Blank copies of Consent Forms including:
    - Student Participant Consent
    - Parent for Student Participant Consent
    - Parent Participant Consent
    - *Please note that while Consent forms include an option for archiving of audio recordings to be released in 2067, these may and/or may not be added to the repository in the future.*
- 6 Blank copies for Plain Language Statements and Coversheet
    - Student Participant Plain Language Statement
    - Student Participant Plain Language Statement Coversheet
    - Parent for Student Plain Language Statement
    - Parent for Student Plain Language Statement Coversheet
    - Parent Participant Plain Language Statement
    - Parent Participant Plain Language Statement Coversheet
- 7 pdf files from makingfutures.net webpages
    - Collingwood
    - Manor Lakes
    - Warrnambool
    - Project
    - Interviews
    - Working Paper
    - Related Publications
- Ethics Application

SIP only files (files not in published dataset)

A total of 134 files are SIP-only files. These include:

- 1 spreadsheet file detailing participant data including biographical information with identifiers
- 70 original non-anonymised interview transcripts with identifiers
- 16 anonymised interview transcripts without archival consent (no identifiers, pseudonyms only)
- 46 signed consent forms with identifiers

Archival process

I began the interview transcription anonymisation process by double-checking each interview transcription file with the corresponding file. During this process, I mainly scanned through each transcription document for identified instances of unclarity. In most cases, I was able to listen to the audio recording and add in the correct words which were mostly educational jargon, names of places and organisations and/or companies. In cases where the audio was too unclear, I left it as is, that is, *e.g.,* [unclear 38:42]. Places and locales identified the interview data have a significant value due to the interests and objectives of the *Making Futures* project. For this reason, I only changed the names of places, spaces and locales if the identity of the participant could be traced back to them. At times when I was uncertain if the identity of participant could be revealed, for example, by their place of their employment, I did my own simple Google check. It should be noted that while I may not have been able to determine their identity through a search engine check, a co-worker may be able to identify the participant by personal anecdotes, history and other biographical details revealed in the interview. It is unlikely like a co-worker, or member of the public will have access to these data files and so it is unlikely that a participant can be identified.

Changes Made

The next step in the process involved anonymisation. I copy-pasted each original interview transcript into a new document and tracked every change. Each change can be identified by the use of [square brackets]. The date each document was anonymised was included in the header of each document. A copy of the original transcript document, the tracked changes document and final anonymised document exists on the project server.

- Each participant's name was replaced with a pseudonym. Pseudonyms were chosen by the Chief Investigator, Professor Julie McLeod and attempted to reflect
- Names of friends, family members, doctors, teachers, mentors and other people were replaced with a description of their relationship to the participant where appropriate; e.g., Joe was changed to [my brother], Mr Joe changed to [my biology teacher] or the initial of their first name (e.g., Joe was changed to [J]).
- The only exception to this was if a parent referred to their son or daughter that has consented to participating in the project, in which case their pseudonym was used; e.g., [Connor].
- Names of places of employment where a participant could be easily traced back to and identified; e.g., a family business which is the only one of its kind in the locale was changed to a descriptive epithet (e.g., [IT business]).

Things that remained unchanged

- Sensitive issues; e.g., self-harm were not censored
- Names of celebrities, politicians, sportspeople
- Names of other schools
- Names of educational and mentoring programs
- Names of streets remained except when a participant's identity could be revealed; e.g., a participant named the street which the business they owned was on was replaced with the first initial [L Street]
- Names of places participants identified as having a significant impact on them. This was usually places participants used to live or visited on holidays.
- Names of places of employment where a participant cannot be easily identified; e.g., franchises or large companies such as banks, supermarkets etc. remained

After each transcript was anonymised in the tracked changes document, it was copy-pasted into a new document which included all changes identified by [square brackets] was then converted to a PDF file, zipped and password encrypted before finally being uploaded to the ADA Dataverse repository.

Reflection – challenges and benefits

The primary challenges throughout this project were associated with overcoming technical issues including a range of issues listed below:

- Troubleshooting with the Terminal app used to zip and encrypt data files
- Keeping track of large volumes of data files generated
- Understanding ADA jargon especially on some pages of the wiki and the licensing forms (see https://docs.ada.edu.au/index.php/Main_Page)

As a graduate research student with no prior knowledge of the research project, one unexpected benefit of having access to research materials and interview data in the capacity I did, was exposure to the interview skills and techniques employed by the researcher. This was beneficial as it gave me insight into how to conduct interviews in an engaging and meaningful manner. Despite the technical challenges I encountered given the nature of this archival work, I enjoyed liaising with ADA experts, my manager and other people self-depositing as it allowed me to gain valuable skills and insight into the future possibilities of qualitative research in Australia.

Appendix 6: Our Lives Asylum Seekers

Project led by Zlatko Skrbiš and Jacqueline Laughland-Booÿ. Summary prepared by Henry Reese and Rachel Flenley

Project summary

The Our Lives Project is a longitudinal study that follows the social and political orientations of a single age cohort of young people in Queensland, as they move from adolescence into adulthood. The project started in 2006, when participants commenced high school at ages 12/13 and aims to check in with participants every two years to chart changes in their life pathways. The study participants are now aged in their mid-twenties. The five stated aims of the project are:

1. To track young people's experiences of major life events, such as tertiary graduation, starting a full-time job, marriage and family formation, leaving the family home, and how these affect their values, behaviours and quality of life in early adulthood.
2. To identify those characteristics of youth transitions which generate positive career, relationship, housing and health outcomes for young people, and those which expose young people to risks of unemployment, tertiary non-completion, residential and relationship instability, and poorer mental and physical wellbeing.
3. To interrogate theoretical notions of 'emerging adulthood', including configurations of structure and agency associated with particular transitional arrangements and how these vary across institutional contexts.
4. To collect new data on a valuable longitudinal cohort, and analyse transitions from secondary schooling in adolescence, towards temporary or more permanent work, family, housing destinations in adulthood.
5. To use innovative social research methodologies, including longitudinal quantitative, qualitative, and mixed methods research to explain varied youth transitions and outcomes.[11]

The principal chief investigator on this study is Zlatko Skrbiš. More information about the overall study can be found here.[12]

The qualitative dataset in question relates to a series of interviews with *Our Lives* participants regarding their views on asylum seekers and 'boat people' in Australia. Jacqueline Laughland-Booÿ's research arose out of Wave 3 of the survey, which started in 2010. At this stage, the survey participants were aged 16/17 and in their final year of secondary education in Queensland. Against a backdrop of heightened public discussion and political concern about asylum seekers, the Wave 3 cohort was asked for the first time to provide their views about 'boat people.' They were asked to respond to a statement that originally appeared in the Australian Election Study (AES) survey, that 'All boats carrying asylum seekers should be turned back.' The Our Lives cohort's responses to this statement were measured on a five-point Likert scale, ranging from 'strongly agree' to 'strongly disagree'. While this question provided useful sociological data about

[11] 'Project Overview,' *Our Lives: The Social Futures and Life Pathways Project*, 2018, https://ourlives.org.au/project-overview/, accessed 4 October 2019.
[12] *Our Lives: The Social Futures and Life Pathways Project*, 2018, http://www.ourlives.org.au/, accessed 4 October 2019.

who among the *Our Lives* cohort might be more accepting of asylum seekers, Jacqueline Laughland-Booÿ wanted to find out *why* they might be more or less accepting.

Using their 2010 survey responses as a guide, in 2012, Jacqueline identified 20 *Our Lives* participants who had firm views on the issue of asylum seekers and conducted a series of qualitative interviews about their views on the matter. Interviewees were conducted with young Queenslanders from a diverse range of socioeconomic backgrounds. All were now in the immediate aftermath of their high school years and had attended a diverse range of schools (independent, Catholic and state) and professed to hold a diverse array of political backgrounds. All participants were born in Australia and said that English was their main language spoken at home. All interviews lasted between 40 and 90 minutes.

The focus of each interview was to expand on the participant's earlier response to the statement that boats carrying asylum seekers should be turned away from Australia. Interviews began by asking the interviewee about what they were currently doing, before turning to whether they recalled their response to the 'boat people' survey item and how they would now rate their response. From this, Jacqueline moved into the reasons for these responses, and the factors the interviewees considered as influences on their opinions. The discussion would then move into broader political issues, including current politicians.

Materials archived

Nineteen interview summaries, representing the 20 young Queenslanders interviewed by Jacqueline-Laughland-Booÿ in 2012, were archived.[13] All participants names were anonymised. These summaries consist of a few paragraphs covering the main points touched on in the interviews, as outlined above. The shortest transcript is one page long and the longest extends to three pages. These interview summaries are *not* full transcripts, although many quotes from the transcripts are used throughout.

In addition to these 19 summaries, I archived several supplementary materials. These included Jacqueline Laughland-Booÿ's 'Research Background' document, providing extensive background to the study itself and its research outcomes, a table providing an at-a-glance summary of the interview summaries, and two published articles that drew on the interview data.

Work required to develop the archive (Rachel Flenley)

The interviews were conducted by Jacqueline Laughland-Booÿ and later transcribed by a third-party transcription service. In the transcriptions, the labels 'interviewer' and 'interviewee' were used to identify the speakers. The files were named using first name pseudonyms developed by Jacqueline Laughland-Booÿ. Rachel Flenley was engaged to develop summaries of these interviews which were then passed to Jacqueline for review. This work included:

- Liaising with Jacqueline on summary structure and depth
- Creating summaries including making de-identification decisions
- Preparing this report

[13] Two interview participants, Katrina and Alice, were interviewed together.

Process/challenges encountered: summary development

Structure

In line with the interview foci the summaries have been organised thematically into three sections: personal circumstances (2012); interviewees' views on asylum seekers and turning back the boats on which they arrived; and their political views and ideas more broadly. This thematic structure means that the summaries do not always or necessarily represent the order of the discussion as it occurred as, at times, the interviews would return to and/or further develop previous points. One other temporal decision of note was choosing a tense for reporting the interviews. After initially deciding to use past tense, I changed to present so as to more easily distinguish between events and ideas of 2012 (the present of the interviews) and those prior to 2012.

De-identification

While Jacqueline had already created pseudonyms for the interviewees, other de-identification protocols had to be employed. These are summarised below, including discussion of some challenging questions that arose as part of this process.

People

Family members, teachers and other school staff, and any other people personally known to and referred to in the interviews have been denoted in generic relational terms. However, names for public figures (e.g., Prime Ministers Howard and Gillard, President Bush etc.) and the political parties to which they belong (e.g., The Greens, Democrats) have been retained, as these cannot lead to identification of the interviewees and offer valuable contextual and historical information. In terms of the interviews themselves, it is worth noting that two sisters asked to be interviewed together. In the transcript, both were labelled 'interviewee' and it was generally not possible to identify the specific speaker. 'They' and 'their' were used in this case. In all cases, quotes have been used throughout the summaries where I felt that changing the words would alter the meaning or sense of the interview unnecessarily.

Place and space

Because southeast Queensland has comparatively few universities and large country towns which might lead to interviewee recognition, I removed specific names and replaced them with generic terms representative of the context (e.g., 'regional centre'). I also removed all school names and replaced them with terms such as 'Catholic girls' school'. Similarly, I anonymised local place names and businesses and used general identifiers such as 'a pizza place' to represent interviewee worksites. International and interstate place names were retained.

Reference to experiences and events

Difficult personal events such as ongoing illness were represented in broad terms. Others, such as travel destinations or living aspirations were recorded as presented by the interviewees, as were references to historical events such as '9/11' and the 'war in Iraq'.

Work required to develop the archive (Henry Reese)

Most of the initial work was conducted by Rachel Flenley, in consultation with Jacqueline Laughland-Booÿ. The materials were passed to me at an advanced phase of the process. All relevant anonymisation and transcribing had already been done. The files were ready to prepare for the ADA Dataverse.

Based on Jacqueline's 'Research Background' document, I created a table outlining the interviewees and their responses to the AES 'asylum seekers' statement.

I then went through and 'cleaned' each interview summary, ensuring consistency of spelling and grammar, as well as formatting issues. The files were well prepared and did not require extensive work. After performing a virus and malware check with Norton Antivirus, I saved each interview summary as a writable PDF. I then named each file in accordance with the ADA's naming conventions, as outlined in the ADA wiki page. Each file was then ready to zip and encrypt. The process I used for this is set out in full in my reflections on the Curriculum Policies project. This how the files appeared at the moment of uploading to the ADA Dataverse, named according to the ADA conventions:

2_OurLivesAsylumSeekers_2012_Ashleigh_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Thomas_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Taylor_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Samuel_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Rory_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Mandy_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Maddie_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Lily_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Kyle_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Kimberly_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Jess_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Jemma_ADAID_.zip
2_OurLivesAsylumSeekers_2012_James_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Emma_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Daniel_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Callum_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Caitie_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Alice_and_Katrina_ADAID_.zip
2_OurLivesAsylumSeekers_2012_Ben_ADAID_.zip

Once uploaded into the ADA Dataverse, I assigned the relevant metadata to each file.

Challenges encountered

Given the advanced state of the data upon my receipt of it, I did not encounter any significant challenges here. My compliments to Jacqueline Laughland-Booÿ and Rachel Flenley for doing such a neat and consistent job with the data.

Appendix 7: Software-based Data Anonymisation

Report prepared by Geordie Zhang

As noted by the ADA (2014), QDR (2013), UKDS (2012–2019b), and others, anonymisation or de-identification of data can be time consuming and expensive, even when planned ahead of time. Such kind of resourcing requirement can slow down or hinder anonymisation at scale for large qualitative data archives. As one possible approach to make anonymisation at scale more viable, we investigated the current state of software-based anonymisation available for qualitative data in the social sciences.

Four software packages were selected and investigated with a sample interview transcript to see the efficacy of the software. These packages are all freely available on the internet (three are open-source, one is closed-source). The software packages were:

1. UK Data Service Text Anonymiser (https://www.ukdataservice.ac.uk/manage-data/legal-ethical/anonymisation/qualitative.aspx).
2. The Irish Qualitative Data Archive provides an open-source anonymisation software written in Java: https://sourceforge.net/projects/datatool/.
3. NLM-Scrubber, a medical text anonymisation software by the U.S. National Library of Medicine: https://scrubber.nlm.nih.gov/annotation/.
4. Netanos, a named entity recognition based anonymiser.

In order to better understand and analyse software-based anonymisation, we considered what needs to happen at a high (process) level as a framework of analysis. From a process perspective, anonymisation consists of the following two steps:

1. The identification of all terms that need to be replaced with pseudonyms to achieve anonymisation
2. The replacement of all identified terms with their respective pseudonyms

For software-based anonymisation to happen, both of the above steps need to occur, whether completed by the software or by a person (or both). During our investigations, we found that the four software we tested all addressed the two steps above, but were substantially different in their approach and extent/accuracy/completeness.

The UK Data Service Text Anonymiser is the simplest anonymiser of the four tested. It is basically a macro in Microsoft Word that highlights all capitalised words and numbers within a text. The heuristic of the software is that capitalised words, which tend to be proper nouns, and numbers, are often parts of the text that need to be anonymised. The highlighting addresses the step of identification (to a somewhat limited extent), and the replacement is all done manually.

In comparison, the Irish Qualitative Data Archive (IQDA) Qualitative Data Anonymizer software uses a list-based method for anonymisation. The pre-processing in the software involves building a name mappings file, which is essentially a table that lists all words that should be anonymised, and the pseudonyms into which to rename the words. The pre-processing is essential to the function of the software and, while the software can automatically anonymise all instances of each

word in the name mappings file, the building of such a file requires significant manual input from the researcher. From the perspective of the anonymisation process, the identification step is fully manual for the IQDA software package, and once the named mappings file is built, the subsequence replacement step is fully automated.

The other two software, NLM-Scrubber and Netanos, use machine learning, a branch of artificial intelligence where the software does not perform set tasks, but 'learns' to deal with the task specified. The branch of machine learning that deals with unstructured text is called natural language processing (NLP). If NLP is applied to anonymisation, the identification step of the process can be approached using a technique called named entity recognition (NER), where the machine learning software model attempts to classify the words in the text into different entities (e.g. persons, organisations, locations). Once the model has completed performing NER on the unstructured text, the software will then replace all identified terms that belong to the entities to be anonymised with their respective pseudonyms. From a software perspective (as demonstrated by the IQDA anonymiser), it is relatively straight forward to replace all identified terms with their pseudonyms (the identification step is far more difficult for software).

A major difference between NLM-Scrubber and Netanos is that the former is closed source, meaning it is not possible to check the validity or the security of the software at the code level. On the other hand, Netanos is open source, and has the greatest potential going forward. Netanos is also convenient in that there is a pre-built user interface (http://netanos.io), with some scope of customising which categories of entities to anonymise.

What should be noted here is that even with NER, at the current state of the art, it's not possible to come close to addressing some of the subtler nuances of term identification in anonymisation as discussed by the contributors in the appendices. However, what makes NLP based anonymisation worth further research and exploration is the speed with which NLP could process large amounts of text. During testing, for a sample text of 6124 words (27,134 non-space characters), it took around 30 to 60 seconds for http://netanos.io to anonymise the sample text. This could be sped up further by using more powerful servers to host the software (e.g. larger virtual machines on research clouds). Whilst the performance of Netanos on the identification step is not perfect, there is sufficient accuracy and speed that this approach is worth further research and development.

In conclusion, we have analysed the process around software-based anonymisation, and investigated the efficacy of four currently and freely available software packages for anonymisation of qualitative data. The most promising approach uses natural language processing (NLP) based term identification, followed by a software automated replacement of all identified terms with their pseudonyms, as demonstrated by Netanos. At the current in time, NLP can only do entity based identification of terms to be anonymised, cannot address some of the subtler nuances of term identification, and is not perfect in its identification of terms based on entities. On the other hand, NLP based anonymisation is extremely fast, and with advances in such technology, may eventually make viable anonymisation of qualitative datasets at scale. One final comment is the value of open source software for research purposes, as without open access to the source code, it is not possible to make detailed investigations into the viability, accuracy and security of a software, nor make custom improvements to the software.

SOCEY.NET