Starting a Conversation about Open Data in Mathematics Education Research

Tracy Logan
University of Canberra
<Tracy.Logan@canberra.edu.au>

This position paper discusses the role of open access research data within mathematics education, a relatively new initiative across the wider research community. International and national policy documents are explored and examples from both the scientific and social science paradigms of mathematical sciences and mathematics education respectively are provided. Within these examples, some of the more well-known concerns associated with making data open and accessible are acknowledged and debated.

This paper is to provide insights into a research mandate that will become increasingly relevant to mathematics education researchers; namely, the obligation to ensure research data and findings are made public. The paper describes the international context, from both policy and practice perspectives, drawing on specific examples from mathematical sciences and mathematics education within Australia and beyond. The intent of the paper is to establish a critical analysis of current practices.

Within Australia, government funding for research is at a crossroads. There is a growing concern that severe cut backs will eventuate over the next few years. For the top scientists and academics this will be problematic as scarce funds will be even harder to secure. For other researchers, it could spell the end of their research programs. Within these politically uncertain times, simmering under the surface is the question, what will research look like in the future? Who and how will research be funded? In conjunction, there is the an increased awareness that more and more research data are being collected and stored, more often than not in digital forms. Universities around Australia (and indeed the world) are increasingly dealing with a data deluge (Borgman, 2012), with the storage, curation and cost issues associated with large data repositories yet to be fully realised. The philosophies behind such repositories are that data are manageable, connected, accessible, and discoverable. In effect, making the data as open as possible for re-use and re-analysis. The paper provides an overview of open research data both internationally and nationally and describes examples from both the scientific paradigm—mathematical sciences; and the social science paradigm—mathematics education. The distinctions are made to highlight the differences between the two paradigms in the advancement of open research data. Some of the concerns regarding social science data being made available via open access are considered.

International and National Research Policy Perspectives

The capacity to retrieve and share research data is not a new phenomenon. In the years 1996-1998, key stakeholders working on the Human Genome Project (HGP) developed the Bermuda Principles. This was a set of principles that stated the sharing of DNA sequencing information developed from the project should be publicly and freely available within 24 hours of being collected. The release of data pre-publication was ground-breaking across most research fields (Contreras, 2011). Indeed, the Bermuda Principles set the scene for other fields of research to consider benefits of releasing data sets, not necessarily pre-publication of results, but certainly in conjunction with publication (see for example the

2003 Report *Sharing Data from Large-scale Biological Research Projects: A System of Tripartite Responsibility*, commonly known as the Fort Lauderdale agreement). In 2007, the Organisation for Economic Co-operation and Development (OECD) (2007) developed a report outlining guidelines and principals for accessing and sharing data produced by government-funded research. They argued that:

> access to research data increases the returns from public investment in this area; reinforces open scientific inquiry; encourages diversity of studies and opinion; promotes new areas of work and enables the exploration of topics not envisioned by the initial investigators (p. 3).

It was from this point on that the international research community's awareness was heightened. Within the United Kingdom and United States, research funding bodies such as the Economic and Social Research Council (ESRC, 2010), the Wellcome Trust (2010) (UK) and the National Science Foundation (NSF, 2010) (USA) have documented policies stating data management plans and provisions for the sharing of data must be submitted with grant applications, that these sections are subject to review and will be influential in the decision to award the funding. The European Union (European Commission, 2013) also identified the need for policies on open access data within its major research and innovation program called Horizon 2020. All publications and data generated through this funding must comply with their guidelines for open access.

From the Australian perspective, the Australian Code for the Responsible Conduct of Research (Australian Government, 2007) was published outlining the principles and practices of researchers and institutions when conducting research. Section 2 in this document outlined management of data and primary materials. In summary, it highlighted the need to retain data for verification purposes and appropriate access for the wider research community. Around the same time, changes started appearing in the Australian Research Council's (ARC) Discovery Project funding rules for 2008 (Australian Government, 2006) where a section was added (1.4.5. Dissemination of research outputs, p. 13) regarding the dissemination of data and outputs:

> The ARC therefore encourages researchers to consider the benefits of depositing their data and any publications arising from a research project in an appropriate subject and/or institutional repository wherever such a repository is available to the researcher(s). If a researcher is not intending to deposit the data from a project in a repository within a six-month period, he/she should include the reasons in the project's Final Report.

This general statement has remained relatively consistent throughout the Discovery Project funding rules since 2008 and presently, for the funding rules for 2016 Discovery Projects, the statements read:

> A11.5.1 All ARC-funded research projects must comply with the ARC Open Access Policy on the dissemination of research findings, which is available at www.arc.gov.au. In accordance with this policy, any publications arising from a Project must be deposited into an open access institutional repository within a twelve month period from the date of publication.

> A11.5.2 Researchers and institutions have an obligation to care for and maintain research data in accordance with the Australian Code for the Responsible Conduct of Research (2007). The ARC considers data management planning an important part of the responsible conduct of research and strongly encourages the depositing of data arising from a Project in an appropriate publically accessible subject and/or institutional repository. (Australian Government, 2014, p. 19)

The ARC Open Access Policy (Australian Government, 2013a) specifically relates to publications being placed in open access repositories. This is mandatory. However, the interesting change is the separation of publications and data, with researchers being

strongly encouraged to deposit data into repositories. This highlights the increased importance placed on the accessibility of research data to the wider community.

In late 2013 (Australian Government, 2013b), the ARC released the Discovery Projects—Instructions to applicants for funding commencing in 2015. This document generally provides advice to applicants on dealing with the relevant systems and explaining what each section of the proposal should contain. For the first time, that document identified that the project description (Part C) is required to have a heading titled *Management of Data*. This stated that all proposals must "outline plans for the management of data produced as a result of the proposed research, including but not limited to storage, access and re-use arrangements" (Australian Government, 2013, p. 15). Through this inclusion, the ARC is effectively making data management and data re-use an assessable component of the proposal, in a similar vein to the UK and USA systems. As Borgman (2012) commented in relation to the NSF policy on data management, " the NSF has accelerated the conversation about data sharing among stakeholders in publicly funded research" (p. 1061). The separation of publications and data in the ARC funding rules and the inclusion of an assessable component related specifically to data management in the proposal emphases the growing awareness from a political perspective that the data generated by public funding is becoming increasingly valuable and needs to be made accessible.

Data Repositories

There are a myriad of data repositories situated globally, with almost every university having some form of searchable digital repository. This does not take into account government funded resources or independent enterprises. Hence, the main priority over the past few years has been the consolidation of, and access, to all the various data repositories. The UK Data Archive (http://www.data-archive.ac.uk/) provides access to social science and humanities data repositories and across Europe and the USA, re3data.org is a registry of data repositories. These registries provide access to a wide variety of data repositories internationally.

Within Australia, since 2004 previous and current federal governments have invested approximately $2.5 billion through the National Collaborative Research Infrastructure Strategy (NCRIS) funding scheme to support the infrastructure required to consolidate and coordinate research across Australia (Lowe, 2015). This has included various aspects of big data collections. Table 1 outlines some of the projects undertaken in relation to the consolidation of data.

This paper will focus on the Australian National Data Service (ANDS) and Research Data Australia as the national registry of research data within Australia.

The main aim of ANDS is to create:

a cohesive national collection of research resources and a richer data environment that will:

- Make better use of Australia's research outputs

- Enable Australian researchers to easily publish, discover, access and use data

- Enable new and more efficient research (ANDS, n.d.).

Among other responsibilities, ANDS developed and currently manages Research Data Australia, a searchable registry of data. This registry provides access to a large number of research data, projects, documents, people, institutions and groups. It has been designed utilising the following categories: Collections; Parties; Activities; and Services. Collections

are research datasets or collections of research materials. Parties are researchers or research organisations that create or maintain research data sets or collections. Activities are projects or programs that create research data sets and collections. Services are the services that support the creation and use of research data sets and collections. Entries are categorised accordingly and there are linking nodes among these categories. With regard to access, there are three levels of access identified within Research Data Australia: Open; Conditional; and Restricted. Open access is defined as online data that can be electronically accessed free of charge with no conditions imposed on the user. Conditional access is seen as online or offline data that can be accessed free of charge, providing certain conditions are met (e.g., registration is required to access data online). Restricted access is online or offline data where access to the data is heavily restricted.

Table 1.

A Sample of Projects Undertaken Through NCRIS Funding to Support Data Consolidation

Projects	Description
National Computing Infrastructure and Supercomputing Centre	High-end supercomputing services to researchers.
Research Data Storage Initiative	Supporting national data storage
National eResearch Collaboration Tools and Resources	Desktop-based data analysis and modelling tools for researchers
Australian National Data Service (including Research Data Australia), National Research Network and Australian Access Federation	Building better electronic communication, connectivity and collaboration networks between national and international research institutions
Australian Data Archive and Australian Data Archive Social Science	Collection and preservation of digital research data

Note: Adapted from Lowe (2015).

The information within Research Data Australia is supposed to represent all fields of research within Australia, so in order to understand how mathematics education is situated, a comparison between a scientific paradigm, mathematical sciences and a social science paradigm, mathematics education is presented.

Open Research Data in Two Paradigms

Within mathematics education, and education more generally, there is an increasing awareness of data storage and re-use. However, compared to the mathematical sciences, education appears to be well behind in their understanding of, and participation in, making research data more open. To demonstrate this, a brief comparison is presented between the scientific paradigm and the social science paradigm. A search was conducted of Research Data Australia to determine the number of entries under mathematical sciences and Education. As described above, entries are represented by collections, parties, activities, or services. The entries are also collated under subjects according to the ANZSRC Field of Research (FoR) classification. It was through these subject classifications that the search was initially conducted. It should be noted that if the entry was not attached to a specific FoR, it does not show up in these classifications, but may be identifiable through other keywords searches. As such, subsequent keyword searches were conducted to identify the

number of collections, parties, activities, and services related to the keywords. These keyword searches also enabled filtering to identify those entries with open data access.

Scientific Paradigm: Mathematical Sciences

The Mathematical Sciences is the *01* classification under the ANZSRC FoR. It includes research areas such as Applied Mathematics, Statistics, and Pure Mathematics. A search at the two-digit level revealed 12,435 entries linked to this FoR. A keyword search of *mathematical sciences* revealed more than 85,000 entries, as categorised in Table 2.

Table 2.
Number of Entries Identified by Keyword Search of "Mathematical Sciences" and Open Data Licence in Research Data Australia by Category

Category	Mathematical Sciences	Open Data Licence
Collections	57,273	19,999
Parties	2,129	—
Activities	25,495	160
Services	120	—

That is a large number of open data licences, so what does that data actually look like. The data in these fields of research are more often than not quantitative and may contain complex systems of numbers and text and spatial information. Generally, this data relates to environmental, biological, or other physical phenomena as opposed to human subjects. It could be argued that much of this type of data is objective and factually based.

Many areas in these sciences have established data archiving and sharing practices, with some academic journals even making it a condition of publication that data be deposited into a publicly accessible database or provided as appendices for others to access (Borgman, 2012). However, this is not the case for the social sciences.

Social Science Paradigm: Mathematics Education

Education is the *13* classification under the ANZSRC FoR and includes Education Systems, Curriculum and Pedagogy, and Specialist Studies in Education. Under the two-digit code, 280 entries are identified. This is an underwhelming amount and there is a large difference in the number of entries between the two subject codes at this level. A keyword search for *mathematics education* revealed 73 entries as categorised in Table 3. None of the entries provided open data licences; however, almost all of the collections indicated an available data set. It is acknowledged that mathematics education is a much more specialised field compared to the general classification of mathematical sciences; however, even at the two-digit level, the differences are stark.

The data sets linked to those collections were classified as conditional or restricted access, which required contacting the chief investigator or the research group/institution to negotiate terms and conditions of use. For example, the research team at the International Centre for Classroom Research at the University of Melbourne have listed all their data sets from the International Learner Perspective Study. However, access must be negotiated with the Centre.

Without an openly available data set to compare with the mathematical sciences, the following section draws on the literature to better understand what mathematics education

data might look like and highlights some of the common issues associated with openly sharing this type of data.

Table 3.
Number of Entries Identified by Keyword Search of "Mathematics Education" and Data Sets in Research Data Australia by Category

Category	Mathematics Education	Data Sets
Collections	45	44
Parties	18	—
Activities	10	—
Services	0	—

Understanding Mathematics Education Data

Mathematics education research data comes in varied forms. Similar to other social science research and depending upon methodology, it can include surveys, interviews, focus groups, tests, classroom observations, policies and other documentation, and various types of digital media such as audio and video recordings. Much of the data collected within mathematics education is rich qualitative data; however, quantitative data is also widely collected. It could be argued that this type of data is subjective insomuch as it specifically relates to human endeavour and behaviour.

There has been much research attention afforded to the storage, archiving and re-use of qualitative data (Bishop, 2012; Cheshire, 2009; Cheshire, Broom, & Emmison, 2009; Corti, 2012; Fielding, 2004; Hammersley, 1997; Mauther & Parry, 2009). Overwhelmingly, the debate revolves around four main areas as identified by Cheshire (2009):

> Broadly, these concerns revolve around issues of research ethics, specifically informed consent and participant confidentiality; data security and access; intellectual property; and the enhanced insight into meaning that is gained from being involved in the data collection enterprise and which is subsequently lost in any secondary analysis. (p. 27)

These four issues will be discussed briefly to highlight the nature of the debate and identify any steps that have been taken to alleviate some of these issues.

Ethics, Security, and Access

The ethical issues with storing and re-using data from human participants tend to focus on the type of informed consent provided at the beginning of data collection and the need to maintain confidentially. Previously, participants were told that after a certain period of time their data would be destroyed and that only members of the research team would have access to it. Hence, the majority of research conducted under those ethics will never be able to be re-used outside of the research team. Those terms have changed and now participants need to be informed about how their data will be kept and that other researchers may have access to the de-identified data. There are real possibilities that participation in research from the Education sector may decline because of these requirements. Certainly when researching sensitive areas, such as different cultures, often the participants only consent because their words, information or data will only be heard or seen by the research team, and often it has taken years of developing trust to get to even

that point (Cheshire, 2009). Coinciding with this is the levels of security and access that others have to the data sets. Much of this can be decided upon by the researcher. As was demonstrated in the example above, many of the mathematics education data sets in Research Data Australian are restricted access, meaning that any form of re-use is negotiated with the owner of the data. ANDS recently published a guide to publishing sensitive data (Olesen, 2014). This outlines some of the steps that can be taken to make sensitive data more open and accessible through data repositories.

Intellectual Property

The majority of research projects that actually get funded are a result of the reputation and knowledge and skills of the chief investigator and the research team. Not only does the idea have to be good and the methodology sound, the researchers must be deemed fit to carry out the project. In some circumstances, the collection of the data comes at a personal cost also. Hence, it is little wonder that many researchers covet their data. However, the data itself actually belong to the researcher's institution and upon retirement or leaving, that data remains the property of that institution.

Context

Research conducted with human participants and about the characteristics of those participants is contextually based. Without context, much of the data is sometimes rendered meaningless and often very hard to interpret. Bishop (2012) identified that "for qualitative methodology, a key issue is context, as data are deemed inseparable from the context in which they are generated" (p. 345). In order to store data and make it appropriate for re-use, often very detailed descriptions of the context of data collection will be required along with data collection instruments and the data itself.

Implications Moving Forward

Given the current political climate and the requirement for ARC funded projects to have their data deposited into a repository, conversations need to begin within the mathematics education community about data storage and open data access. The relatively low number of mathematics education entries into Research Data Australia may be indicative of the culture of our research environment, but it may also highlight the difficulty of having a data set that can be easily stored and made accessible. Despite the advances in technology that have allowed such data repositories to exist and function, it could be the case that much of the data collected in mathematics education is done so in non-digital form and hence time, money and equipment are needed to make it repository ready. Alternatively, it could be the case that consent for such storage and access was not sought or not granted by the participants. Regardless of the reasons, research funding is limited and looking into the future, data repositories may be the only viable source of data available to conduct research.

References

Australian Government. (2006). *Australian Research Council: Discovery projects funding rules for funding commencing in 2008*. Retrieved 18 February 2015 from: http://www.arc.gov.au/ncgp/dp/dp_fundingrules.htm

Australian Government. (2007). *Australian Code for the Responsible Conduct of Research*. Retrieved 18 February 2015 from https://www.nhmrc.gov.au/guidelines-publications/r39

Australian Government. (2013a). *ARC open access policy*. Retrieved 18 February 2015 from http://www.arc.gov.au/applicants/open_access.htm

Australian Government. (2013b). *ARC Discovery Projects—Instructions to applicants for funding commencing in 2015*. Retrieved 17 March 2015 from http://www.arc.gov.au/pdf/DP15/DP15_ITA.pdf

Australian Government. (2014). *Australian Research Council: Funding Rules for schemes under the Discovery Program for the years 2015 and 2016—Australian Laureate Fellowships, Discovery Projects, Discovery Early Career Researcher Award and Discovery Indigenous*. Retrieved 18 February 2015 from: http://www.arc.gov.au/ncgp/dp/dp_fundingrules.htm

Australian National Data Service [ANDS] (n.d.). About ANDS. Retrieved 17 March 2015 from http://ands.org.au/about-ands.html

Bishop, L. (2012). Using archived qualitative data for teaching: Practical and ethical considerations, *International Journal of Social Research Methodology*, *15*(4), 341-350. http://dx.doi.org/10.1080/13645579.2012.688335

Borgman, C.L. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology*, *63*(6), 1059–1078.

Cheshire, L. (2009). *Archiving qualitative data: Prospects and challenges of data preservation and sharing among Australian qualitative researchers*, Discussion paper of The Australian Qualitative Archive, University of Queensland, Brisbane, Australia. Retrieved 17 March 2015 from http://www.assda.edu.au/forms/AQuAQualitativeArchiving_DiscussionPaper_FinalNov09.pdf

Cheshire, L., Broom, A., & Emmison, M. (2009). Archiving qualitative data in Australia – an introduction. *Australian Journal of Social Issues*, *44*(3), 239-254.

Contreras, J.L. (August, 2010). Bermuda's legacy: Policy, patents and the design of the genome commons. Washington University in St. Louis Legal Studies Research Paper No. 10-09-02. Retrieved 17 March 2015 from: http://ssrn.com/abstract=1667659

Corti, L. (2012). Recent developments in archiving social research. *International Journal of Social Research Methodology*, *15*(4), 281-290. http://dx.doi.org/10.1080/13645579.2012.688310

Economic and Social Research Council. (2010). ESRC Research Data Policy (updated 2013). Retrieved 17 March 2015 from http://www.esrc.ac.uk/_images/Research_Data_Policy_2010_tcm8-4595.pdf

European Commission. (2013). *The EU framework programme for research and innovation: Horizon 2020. Guidelines on open access to scientific publications and research data in Horizon 2020*. Retrieved 17 March 2015 from http://ec.europa.eu/research/swafs/index.cfm?pg=policy&lib=science

Fielding, N. (2004). Getting the most from archived qualitative data: Epistemological, practical and professional obstacles. *International Journal of Social Research Methodology*, *7*(1), 97-104. http://dx.doi.org/10.1080/13645570310001640699

Hammersley, M. (1997). Qualitative data archiving: Some reflections on its prospects and problems. *Sociology*, *31*, 131-142.

Knight, S., Nguyen, H.X., Falkner, N., Bowden, R. & Roughan, M. (2011). The Internet topology zoo. *IEEE Journal on Selected Areas in Communications*, *29*(9), 1765-1775. Retrieved 18 March 2015 from http://www.topology-zoo.org/gallery.html

Lowe, A. (2015). Explainer: the National Collaborative Research Infrastructure Strategy (NCRIS). The Conversation Australia. Retrieved 18 March 2015 from http://theconversation.com/explainer-the-national-collaborative-research-infrastructure-strategy-ncris-38837

Mauther, N.S., & Parry, O. (2009). Qualitative data preservation and sharing in the social sciences: On whose philosophical terms? *Australian Journal of Social Issues*, *44*(3), 291-307.

Olesen, S. (2014). *ANDS guide: Publishing and sharing sensitive data*. Canberra: Australian National Data Service.

Organisation for Economic Co-operation and Development (OECD). (2007). *OECD principles and guidelines for access to research data from public funding*. Paris, OECD. Retrieved 18 February 2015 from http://www.oecd.org/science/scitech/oecdprinciplesandguidelinesforaccesstoresearchdatafrompublicfunding.htm

National Science Foundation. (2010). NSF Data Sharing Policy. Retrieved 17 March 2015 from http://www.nsf.gov/bfa/dias/policy/dmp.jsp

Wellcome Trust. (2010). Policy on data management and sharing. Retrieved 17 March 2015 from http://www.wellcome.ac.uk/About-us/Policy/Policy-and-position-statements/WTX035043.htm