IJDC | *General Article*

Curation After the Fact: Practical and Ethical Challenges of Archiving Legacy Evaluation Data

Ruth Geraghty
Children's Research Network of Ireland and Northern Ireland

Abstract

Over a 12-year period, the Atlantic Philanthropies invested more than €127m in agencies and community groups, running 52 prevention and early intervention (PEI) programmes and services in the children and youth sector throughout Ireland. As a condition of this funding, each PEI programme was evaluated by a university-based research team, resulting in a substantial collection of metric and qualitative information about ways to improve the lives of vulnerable Irish families. In 2016, the Atlantic Philanthropies funded the Prevention and Early Intervention Research Initiative at the Children's Research Network of Ireland and Northern Ireland (hereafter, the Initiative) to gather, prepare and share this evaluation data through the public data archives.

The Initiative faces several challenges in its objective to archive this extensive collection of legacy data, and this paper will present two of the more salient challenges: how to share this data so that it is both (1) meaningful and (2) ethical. The paper pays particular attention to the challenges of safely sharing evaluation data through anonymisation and restricted access conditions; and also, the practical and ethical challenges of retroactively preparing these datasets for the archive.

A series of publicly available documents that guide each stage of the Initiative are in development, and are emerging as a key output. This paper will describe two pivotal documents, namely the CRN-PEI Guiding Principles, and the CRN-PEI Protocols for preparing and archiving evaluation data. The CRN-PEI Guiding Principles outline the key legal and ethical obligations of archiving this legacy evaluation data, and act as moral compass to steer our progress through these uncharted waters. The CRN-PEI Protocols define the standards for how data included in the Initiative is prepared for deposition in the public data archives, so they are easily located, interpretable and comparable in the long term. This protocol is based upon best practice documentation from a number of international sources and our primary aim is to generate 'safe, useful data' (Elliot at al., 2016).

Received 23 February 2017 ~ *Accepted* 8 December 2017

Correspondence should be addressed to Ruth Geraghty, Children Research Network for Ireland and Northern Ireland, Centre for Effective Services, 9 Harcourt Street, Dublin 2, Ireland. Email: rgeraghty.crn@effectiveservices.org

An earlier version of this paper was presented at the 12th International Digital Conference.

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

This paper describes a project that is currently underway at the Children's Research Network of Ireland and Northern Ireland, to archive social research data that was generated by a series of evaluations from 2004 to 2016. A major element of the work is to clean, organise and anonymise the content of these data so that they can be safely and meaningful re-used by a wide community of academic users, for further scientific research. The data are of value not merely for scientific purposes, but also because this research was unprecedented internationally in its contribution to the development of evidence-based prevention and early intervention (PEI) services for children and young adults. The term 'legacy' is therefore used here to reflect the exceptional value of the data in capturing a key period in PEI development. Additionally, the data is of value beyond the field of evaluation, as many of these collections captured important social, political and economic information during a key phase of Irish life. The term 'legacy' also alludes to the fact that this data was created at a point in time when it was not yet common practice in Irish social research to consider archiving during research planning.

The Story Behind the Data (2004 – 2016)

Over a 12-year period, the Atlantic Philanthropies made significant investment into the development and delivery of community based PEI services in the children and youth sector in the Republic of Ireland and Northern Ireland, and a condition of this funding was independent evaluation of each programme. Researchers from several universities in both the Republic and Northern Ireland, plus a small number of independent research agencies, were recruited to conduct these evaluations. Evaluation was typically two-fold and involved (1) an 'outcomes evaluation' using a statistical survey to measure the impact on the population receiving the PEI service, and (2) a 'process evaluation' using mainly qualitative interview methods to capture the opinions and practices of different stakeholder groups involved in the PEI service. In most cases data collection was carried out over two or three years, although one evaluation was longitudinal and data was collected over five years.

These PEI evaluations were predominantly experimental, and followed a Randomised Control Trial or quasi-experimental research design. In experimental research, information is captured about the study population at a series of time points, in order to measure the before-and-after effects of an intervention. For example, the PEI evaluations usually surveyed participants in receipt of a programme before the programme started (the baseline data), in the middle of the programme delivery (the mid-phase data) and at completion of the programme (the end-phase data). The baseline data is the most sensitive amongst these files, in terms of risk of identification, as it tends to contain the largest quantity of demographic characteristics about each participant. The combination of demographic variables are typically used during the analysis to assess causal factors for the standardised scores that make up the remainder of the dataset.

In this research contract, ownership of all data created during an evaluation belonged to the PEI organisation. However, in practice, it was common for the research team to hold on to the data at the site of their university or research company, and to

 doi:10.2218/ijdc.v12i1.550

deliver only the reports from the analysis to the PEI organisation. Many of the PEI organisations did not have internal research staff, and did not pursue the data files on completion of the evaluation. Additionally, the researchers that conducted the evaluation submitted their research proposal to their university for ethical review, and complied with the university's rules for data collection and storage. Ethical guidance from the university frequently included the conditions that the data would be held securely and exclusively on the site of the university, and would only be visible to members of the university research team. There was no definitive long-term plan for the data, nor a plan for its transfer back to the PEI organisation on completion of the research. In some instances, this ethical guidance contained a vague statement about the destruction of 'personal data' after a certain period, without clarifying which elements of the project were to be treated as personal data.

The PEI Research Initiative (2016 – 2017)

Given the scale of investment and the fact that much of the data was never fully mined, in early 2016, the Atlantic Philanthropies funded an archiving project at the Children's Research Network in Dublin, entitled the 'PEI Research Initiative' (hereafter, the Initiative). Now in its second year, the Initiative is tasked with archiving a substantial portion of the total number of datasets created during the 12-year period of evaluation, in collaboration with the data-owning PEI organisations and their academic evaluators. This archiving project is pioneering in Ireland as it is the first attempt to preserve, and share, a large amount of evaluation data from a range of sources. The Initiative also promotes a culture of data re-use by disseminating a series of grants for secondary analysis of the datasets that are archived. The intention is to maximise the original investment by generating new knowledge about PEI, but the project also has a knock-on effect of introducing a new cohort of researchers, working in evaluation research, to archiving and secondary analysis of archival sources.

Individual evaluations were led by university-based teams from a range of third-level institutions across Ireland and the UK, and mainly from the disciplinary areas of education, economics, sociology, nursing and psychology. Although the majority of the evaluations followed an experimental design, some disciplinary effects are evident, for example the way in which missing data was managed, using a dot in some cases and a coded numeric value in others. Each collection of project files is distinct in terms of the data management and file naming protocols that were followed (or not) by each research team, the software formats favoured by individual researchers, and the availability and quality of contextual material. The first challenge of the Initiative is, therefore, to align a disparate group of datasets into a standardised format so they can be meaningfully brought into dialogue with one another, without losing what is uniquely valuable within each.

At a minimum, data from the outcome evaluations will be made available through the Irish Social Science Data Archive[1] and the UK Data Archive[2], both of which are public data archives with the appropriate access conditions for sensitive, quantitative data. Where possible, data from the process evaluations will be made available through the Irish Qualitative Data Archive[3] and the UK Data Archive, both of which have the

1 Irish Social Science Data Archive: https://www.ucd.ie/issda/
2 UK Data Archive: http://www.data-archive.ac.uk/
3 Irish Qualitative Data Archive: https://www.maynoothuniversity.ie/iqda

necessary infrastructure for sensitive qualitative data. Once a PEI organisation is willing to share their data, the Initiative's Data Curator works closely with the researchers to asses which parts of the evaluation are suitable for processing, and together they develop a data processing plan. Processing occurs on the site of the university, and this work consists of cleaning and anonymising the data. Curation is carried out at the Children's Research Network and includes writing collection-level metadata (per evaluation) and the preparation of contextual materials including user guides, data dictionaries and copies of data collection tools, participant information and consent forms.

The majority of the organisations and researchers have been enthusiastic about archiving their data through the Initiative, as they are cognisant of the value of safely storing it for posterity, as well as the opportunity this creates to reuse it alongside data from other evaluations. The datasets to be included are in various conditions and sometimes spread across a range of locations, and researchers may not have engaged with these files for many years. For most of the researchers, this is their first experience of depositing data with a public data archive. To assist researchers in revisiting these 'old' files, the Initiative provides financial support through a series of archiving grants, and a dedicated Data Curator to provide technical support, training and a liaison between the research team and the archives. These two provisions have been fundamental to researcher participation in the Initiative. However, when plans for archiving are made at the end of a research project or, as in our case, after the project has completed, archiving the data is a much more difficult task, as one must retrofit quality control and anonymisation protocols on the data after the fact. Two of the more salient challenges for the Initiative are considered here, namely how to share this legacy data so that it is (1) meaningful and (2) ethical.

The strategies presented here are captured in two documents (currently in preparation) that are pivotal to our work, namely, (1) the CRN Guiding Principles which outline the key legal and ethical obligations of archiving legacy evaluation data, and act as moral compass to steer our progress through these uncharted waters, and (2) the CRN Protocols for Preparing and Archiving Evaluation Data which define the standards for how data included in the Initiative are prepared for deposition in the public data archives, so they are discoverable, interpretable and interoperable. As our primary aim is to support researchers to process their data, these guidance documents are researcher-orientated, and based upon best practice in archiving social science data from sources in Europe and the USA. Although these documents address the archiving of evaluation data, much of the guidance can be applied to other research paradigms and we have been requested to share these documents with various research teams that are planning to archive their own (non-evaluation) data. We have therefore discovered, quite by accident, that the documents fill a gap amongst the wider Irish research community.

Evaluation research is an extreme example of the sensitivities and tensions that are intrinsic to social research, as at the very core of this methodology is assessment or measurement of people, organisations or services. There are multiple stakeholders in evaluation research, and the stakes are high if the resulting data is not carefully managed. Examples of such include the measurement of a child's educational development in comparison with her classmates using standardised scales; or the feedback from an employee on their place of work during a confidential interview. In contrast to this, there is an ethos of open knowledge and sharing in the PEI community, which is made up of voluntary and community organisations. A good amount of information about the PEI initiatives is available in the public domain, for example, most of the research findings exist as openly available reports on the websites of the

 doi:10.2218/ijdc.v12i1.550

commissioning organisations, and in many cases include a list of urban and suburban areas where the intervention was rolled out. This raises significant challenges to the Initiative for anonymising the research data.

The Initiative therefore offers a case study of maintaining individual confidentiality while providing access to the widest possible audience of 'genuine' users, in the context of retrospectively applying best archiving practices on this data. At the time of writing this paper, these issues were in no way fully resolved, nonetheless the Initiative offers an interesting case for consideration.

Ethics and Archiving Social Science Data

Ethical research with human subjects is founded on fully-informed consent, whereby a participant freely gives their permission to participate in the research on the basis that the researcher has provided clear and comprehensive information on how the data will be collected, used and stored. During this consent process, the researcher will usually make commitments to the participant that their data will be handled in such a way as not to expose them to any harm. This is a standard that has been applied in social research for a long time, however the inclusion of a clause to archive the research data is something new in the consent process. As of early 2017, it is not yet standard practice amongst the social research community to include a provision for archiving data in their application for ethical approval, or in their participant consent form. Changes in funder requirements for open data are likely to have some impact here, for example the extension of the Open Research Data Pilot to include all projects that receive Horizon 2020 funding from 2017 has recently generated a lot of discussion amongst the social and health science research communities on how to actually achieve this.

Guidance from the archives for producing safe and ethical data tends to include the following three conditions: "when gaining informed consent, include provision for data sharing; where needed, protect people's identities by anonymising data; consider controlling access to data" (Van den Eynden et al., 2011). A version of the three-step approach of consent, anonymisation and controlled access is replicated in the advice across the social science archives (see for example CESSDA, ANDS, UKDA and ICPSR) and is currently proposed as the most robust approach to safely sharing data. The archives advise that, at the very minimum, the language used in a consent form should never preclude the possibility of archiving the data in the future, for example "through restrictive language … stating that the data will only be shared by the research team or only in publications"[4] or by "promising to destroy data unnecessarily"[5] or by "promising that the data will only be shared in aggregate form or statistical tables" (ICPSR, 2012).

In the PEI evaluations, very restrictive language was used in the consent forms and/or applications for ethical approval, and is a major obstruction to the work of the Initiative. The evaluations were conducted at a time when it was not yet common practice in Irish social research to consider archiving during research planning. Indeed, the two social science data archives in Ireland were newly founded during the mid-2000s, and researchers were slow to engage with them during this early phase due to a lack of knowledge about the feasibility of data re-use (Geraghty, 2014). The researchers

4 Consortium of European Social Science Data Archives (CESSDA) User Guide on research data management: https://www.cessda.eu/Research-Infrastructure/Training/Research-Data-Management

5 Australian National Data Service Ethics, Consent and Data Sharing Guide: http://www.ands.org.au/guides/ethics-consent-and-data-sharing

were not mandated by their commissioners to archive the evaluation data, nor did the researchers make any provision for archiving, as they were likely unaware of the huge potential for re-use in the long term. Consequently, in most – but not all – cases, research participants were never informed of any plan to share the anonymised data, nor asked to consent to this. This has raised an ethical question of whether archiving this data contravenes the commitments about respondent confidentiality that were made to research participants.

The participants consent forms followed a standard format that is currently promoted by ethics review panels as best practice, yet is very problematic for archiving. For example, the following excerpt is from the parent consent form of Evaluation A: "All information will be held in a locked cabinet at the researchers' place of work and will be accessed only by the research team; no information will be distributed to any other unauthorised individual." In some cases, the commitments made during ethical approval are the source of the problem, for example the language from the ethics application for Evaluation B: "only researchers on the [Evaluation B] team will have access to research material." On the other end of the spectrum, Evaluation C designed their consent form in consultation with staff from a data archive from early on in the project, and consequently their consent form includes an explicit reference to archiving:

> 'Once the programme ends an anonymised dataset (with your name and contact details removed) will be placed in the Irish Social Science Data Archive in UCD and may be used by other researchers. Again, this dataset will not contain any of your personal details and names will be replaced by numbers so that any researcher will not be able to identify the responses of any participant' (excerpt from consent form of Evaluation C).

The language used by Evaluation C has certainly made data archiving more workable, however there are still considerable challenges to safely sharing this data using anonymisation.

Addressing the Limitations of Anonymisation

Anonymisation is currently endorsed in national and European law as the optimal procedure for storing and reusing data as "[d]ata which has been irreversibly anonymised ceases to be 'personal data'", and so can be retained and used without having to comply with the Data Protection Acts[6]. This advice from the independent advisory body on data protection in Ireland echoes the incoming General Data Protection Regulation (GDPR), which will be transposed into national law across the EU member states in 2018. Recital 26 of the GDPR specifically addresses the anonymisation of data, whereby "[t]he principles of data protection should … not apply to … personal data rendered anonymous in such a manner that the data subject is not or no longer identifiable"[7].

6 Data Protection Commissioner Anonymisation and Pseudonymisation: https://dataprotection.ie/viewdoc.asp?DocID=1594&ad=1

7 Regulation (EU) 2016/679 of the European Parliament and of the Council 27 April 2016: http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf

 doi:10.2218/ijdc.v12i1.550

Similarly in UK, "[d]ata protection law does not apply to data rendered anonymous in such a way that the data subject is no longer identifiable[8]. So far, the legislation on data protection is quite helpful to a researcher who wishes to store or share their anonymised data indefinitely - whether local research ethics allows for this is a separate issue.

And yet, the special exclusion for anonymised data hinges on the principle of 'irrevocable anonymisation' or data that is 'no longer identifiable'. Such data is exempt from data protection legislation so long as it is persistently anonymous. However the condition of irrevocable anonymisation is harder to achieve in an increasingly connected, digital environment. In the most recent iteration of guidelines on anonymisation, the Irish Data Protection Commissioner recognises that the jury is still out on the effectiveness of various statistically driven disclosure control techniques, which, until recently, were considered the most robust way to ensure data confidentiality. The central problem for anonymisation today is 'reidentification through linkage' (Elliot et al., 2016) where an anonymised file is linked to outside information through a shared piece of data. If the outside information contains an identifier (such as personal names or identification numbers) then the anonymisation is undone. Linkage can be used to link lots of files together so that one file containing identifiers can be linked through a chain of connections to a file containing sensitive information. According to Ohm, "the robust anonymization assumption is deeply flawed" (2010) as any piece of information, however innocuous, can be rendered as 'personal data' if it provides a link between files that eventually leads to re-identification.

Therefore, when talk of 'anonymising' data, what this really means is the removal of the more obvious identification risks in order to make it difficult to re-identify a person – but not impossible. The term 'data that has been processed for anonymisation' is often suggested in the literature as more preferable to 'anonymised data', as it indicates that re-identification has been reduced to an acceptable level, but not entirely removed. This impermanent state is mirrored in the advice on anonymisation from the Data Protection Commissioner:

> 'Organisations don't have to be able to prove that it is impossible for any data subject to be identified in order for an anonymisation technique to be considered successful. Rather, if it can be shown that it is unlikely that a data subject will be identified given the circumstances of the individual case and the state of technology, the data can be considered anonymous'[9].

The objective of anonymisation is to produce data this is safe to share, but also useful and worthwhile, and Elliot et al. show that there is a "trade-off between the two" (2016). In order to strike the best possible balance, they suggest due consideration of the use case of the data – in other words, the data should be anonymised with consideration of who the authentic user will be and what they will most likely want to get from the data. In the Initiative, our interpretation of this advice is to alter the demographic variables which run a higher risk of disclosure, while leaving the standardised score variables untouched. The demographic variables are easier to re-categorise without losing too much of their interpretive power. The score variables, on the other hand, are

8 Information Commissioners Office – Anonymisation Code of Practice: https://ico.org.uk/media/1061/anonymisation-code.pdf

9 Data Protection Commissioner Anonymisation and Pseudonymisation: https://dataprotection.ie/viewdoc.asp?DocID=1594&ad=1

virtually unusable if altered. The final step in our anonymisation process is to change the participant identifier number (the ID variable) to break the primary link between the archived file and any alternative versions of this file that could resurface in the future.

It is also preferable to apply such changes to the data along with some control of the environment into which the data are released. Elliot et al. (2016), described this as an "environment-based solution" by controlling who can access the data, how they can use it and how and where the data is accessed from. Once processed, the evaluation data will be physically located at one archive as a restricted access collection, and will be limited to a defined user group. The archive will attach a permanent identifier to the collection of data from each evaluation, such as a DOI. Openly available, collection-level metadata will exist across all three archives for maximum exposure to potential new users and this will include the DOI information. Those interested in accessing the data must apply to the archive for access and once accredited, they will receive a copy of the processed data through an encrypted delivery channel.

By releasing data through the safeguarded mechanism of the archive, a new user must agree to the specifics of the End User Licence (EUL) that is attached to the data, before they can receive anything. The EUL is not a cure-all as there is always some risk that data files might be shared by accredited users with non-accredited users, misplaced or stolen, but this certainly limits the risk of data being targeted for ill intent, because it is not easy to get to. The EUL also plays an important role in fostering a culture of safe handling of all research data amongst the research community, as the user must consider the duty of care they have to the data they have been entrusted with. For example, the requirement of some EULs that the data is stored in an encrypted location, has encouraged many researchers to investigate encryption software options for their own desktop work for the first time.

Other Work by the Initiative

There are significant ethical questions still to be ironed out in regard to archiving a series of data that is of great value to the public good, but missing participant consent to do so. To further this aspect of the project, we are undertaking an experiment with retrospective consent for qualitative evaluations which had small population samples of twenty participants or less. There are examples of projects which had good success with re-negotiating participant consent to archive data retrospectively (see for example Corti et al., 2000; and Kuula, 2010/2011). We have designed a letter of re-contact with the original research team under the guidance of the staff at the UK Data Archive and the Finnish Data Archive, both of whom have experience in this area. Once the proposal to re-contact the participants receives ethical approval from the university (it is in and around six years since they participated in the study), the letter will be sent from the original research team to the research participants. The letter explains, in accessible language, the benefit for archiving the data, how confidentiality will be managed, where the data will be archived and the type of person that will have access it. The letter also advises the recipients that a follow-up phone call from the original research team will be made to discuss the proposal to archive, but that their consent will be entirely voluntary. Our intention is that these phone calls will capture an indication, both positive and negative, of public sentiment on the archiving project, which might in turn influence our ethical discussion on how to manage all of the PEI legacy data.

 doi:10.2218/ijdc.v12i1.550

Conclusion

Because data archiving and data re-use are relatively new in Irish social research, the Initiative has made great headway in introducing a conversation on ethics, safe data handling and long-term preservation of valuable data amongst the Irish research community, culminating in a roundtable discussion on the topic in Dublin in November 2017[10]. We have brought the issue of what to do with legacy data to the attention of a number of university ethics review committees around Ireland and it is positive that these committees are discussing the issue simultaneously. Legacy data is an issue that is unlikely to disappear from the university campus any time soon, and what we need is joined-up thinking between these independent committees on how to respond to this data. Also, in the current climate of data privacy and security risk, it is good timing to start an honest conversation about what we are doing with research data when a project ends.

Perhaps unintentionally, the Initiative has become a project about building trust amongst the research community, their funders and the research participants. The current ethical grey area of archiving social science data without explicit consent, is problematic, both for depositors and for potential new users of the data, as pointed out by Gonçalves Curty: "Social scientists might hesitate to re-use existing data generated by other researchers if they perceive risks associated with the consent and approval for conducting the study, which was granted only to the original data collectors" (2016). The Initiative highlights the very urgent need for explicit ethical guidance on participant consent to archive, for projects that are happening now and about to happen soon.

Yet, "consent alone does not absolve the responsibility of researchers to anticipate and guard against potential harmful consequences for participants" (Corti at al., 2000). As private citizens, we can be agreeable to terms and conditions without necessarily considering in full their implications, as illustrated by our ease with signing up to the conditions of using social media sites that sell on our data. Even with fully informed consent, it is ultimately the responsibility of the researcher to ensure the data are as secure as possible. The Initiative is contributing to a culture change in how researchers handle data during and after its production. An unintentional output from this project is guidance on how future research data can be archived and safely shared.

The Initiative is experimental: while we cannot present definitive solutions to the challenges raised in this paper, the Initiative has given us a space to try out some techniques to address these challenges. Time for experimenting in this area is a luxury most research projects don't have, so we are keen to share our learning with the research community. This, we believe, will address a skills gap amongst the community as widely as possible, and this project is already disseminating a series of training materials and workshops via the Children's Research Network, which is a membership organisation that is composed of the very community that create and use this data.

References

Elliot, M., Mackey, E., O'Hara, K., & Tudor, C., (2016). *The anonymisation decision-making framework*. Manchester: UKAN University of Manchester.

10 Children's Research Network of Ireland and Northern Ireland annual conference 2017: https://childrensresearchnetwork.org/activity/events/conference-2017

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 Sozialforschung/Forum: Qualitative Social Research*, *1*(3). Retrieved from http://nbn-resolving.de/urn:nbn:de:0114-fqs000372

Geraghty, R. (2014). Attitudes to qualitative archiving in Ireland: Findings from a consultation with the Irish social science community. *Studia Socjologiczne*, 3. ISSN 0039−337

Gonçalves C.R. (2016). Factors influencing research data reuse in the social sciences: An exploratory study. *International Journal of Digital Curation,* 11(1), 96–117. doi:10.2218/ijdc.v11i1.401

Inter-university Consortium for Political and Social Research (ICPSR). (2012). *Guide to social science data preparation and archiving: Best practice throughout the data life cycle* (5th ed.). Ann Arbor, MI. ISBN 978-0-89138-800-5

Kuula, A. (2010/2011). Methodological and ethical dilemmas of archiving qualitative data. *IASSIST Quarterly*, *34*(3&4), 35(1&2), 12-17. ISSN: 0739-113

Ohm, P. (2010). Broken promises of privacy: Responding to the surprising failure of anonymization. *UCLA Law Review*, 57, 1701- 1777. SSRN: https://ssrn.com/abstract=1450006

Van den Eynden, V., Corti, L., Woolard, M., Bishop, L. & Horton, L. (2011). *Managing and sharing data*. Manchester: UKAN University of Manchester.