NIRSA Working Paper Series No 65 – January 2011

NIRSA
National Institute For Regional and Spatial Analysis
An Institiúid Náisiúnta um Anailís Réigiúnach agus Spásúil

NUI MAYNOOTH
Ollscoil na hÉireann Má Nuad

Best Practice in Archiving Qualitative Data

Jane Gray,
Julius Komolafe,
Hazel O'Byrne,
Aileen O'Carroll (Irish Qualitative Data Archive),
Tara Murphy (Tallaght West Childhood Development Initiative)

John Hume Building, National University of Ireland, Maynooth,
Maynooth, Co Kildare, Ireland.
Áras John Hume, Ollscoil na hÉireann, Má Nuad,
Má Nuad, Co Chill Dara, Éire.
Tel: + 353 (0) 1 708 3350 Fax: + 353 (0) 1 7086456
Email: nirsa@nuim.ie Web: http://www.nuim.ie/nirsa

HEA
Higher Education Authority
An tÚdarás um Ard-Oideachas
Funded under the
Programme for Research
in Third Level Institutions (PRTLI),
administered by the HEA

NDP
NATIONAL DEVELOPMENT PLAN

Best Practice in Archiving Qualitative Data

Table of Contents

Best Practice in Archiving Qualitative Data

The value of data archiving is increasingly recognized in Irish research policy. For example, IRCHSS requires that "whenever data is to be collected with the support of a grant awarded by IRCHSS, applicants must specify the means by which that data will be made available as a public good for use by other researchers." The Higher Education Authority policy on open access states that: "Data in general should as far as is feasible be made openly accessible, in keeping with best practice for reproducibility of scientific results."[1] In order to meet these requirements, researchers and research organisations must consider how to manage and process data for archiving, as well as the ethical and legal issues (see Box 1) involved, when planning research projects.

Box 1: Legal and Ethical Framework for Archiving

1. Irish Data Protection Act 1988 and (Amendment) Act 2003. This lays out the requirements that data controllers must meet in order to assure the privacy of individuals.

2. Freedom of Information Act 1997 (FOI) and (Amendment) Act 2003. This obliges government departments, the Health Service Executive (HSE), local authorities and a range of other statutory agencies to publish information on their activities and to make personal information available to citizens.

Social science researchers are also governed by codes of professional ethics that specify their obligations to research informants.

The guidelines of the Sociological Association of Ireland are available at http://www.sociology.ie/

Researchers must also ensure that the costs of archiving data are built into any funding proposal. This handbook aims to assist social science researchers who are planning to archive **qualitative** data. It provides a guide to the following six aspects of best practice in qualitative data archiving:

- ✓ Data management planning
- ✓ Tools and software
- ✓ Data preparation
- ✓ Data processing

[1] http://www.hea.ie/files/files/file/Open%20Access%20pdf_.pdf

- ✓ Data deposit
- ✓ Dissemination of data

These guidelines will be useful both to the researcher who intends to deposit data and to the data curator who maintains the archive. In some cases this may be the same person - for example when the researcher is maintaining a personal data collection. Every qualitative project is different, and it is not possible to produce a template that will address every archiving issue. This handbook is intended to introduce the researcher and data curator to the issues they need to consider when developing an archiving strategy appropriate to their data set.

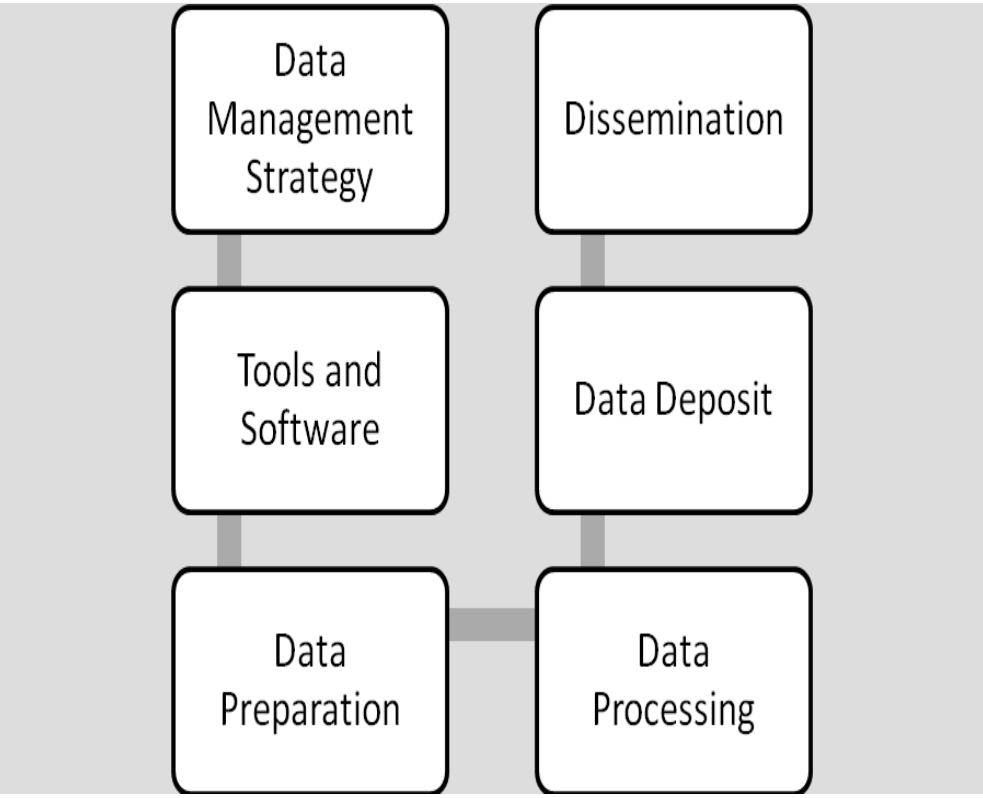
The following guidelines apply principally to textual data that have been generated in qualitative research projects. Where we refer to a 'document' we mean research data in text format. Qualitative social science research generates a range of different types of documents. The data to be archived might include transcripts of interviews, observational notes, or (as was the case in the RACcER[2] project) minutes of meetings. Management and processing of other forms of qualitative data (such as audio or image data) require additional guidelines and the Irish Qualitative Data Archive can offer further advice on this or on any other issues that may be unique to a specific research project.

This handbook on best practice was developed as part of the RACcER research project, a joint initiative of the Irish Qualitative Data Archive and the Tallaght West Childhood Development Initiative (CDI), co-funded by CDI and IRCHSS as part of its Research Development Initiative. The goal of the project was to archive documents associated with CDI and, in the process, to establish best practices in qualitative data archiving (see Box 2). The following guidelines were developed in consultation with the researchers and funders who kindly agreed to be interviewed about their understandings, perceptions and expectations of qualitative data archiving projects[3]. We thank them sincerely for their contribution.

[2] Re-use and Archiving of Complex Community-Based Evaluation Research

[3] We would also like to acknowledge the invaluable advice of Dr. Libby Bishop of Timescapes and the UK Data Archive, and Dr. Arja Kuula of the Finnish Social Science Data Archive. Thanks also to Justina Senkus for her help in preparing this handbook.

Best practice in qualitative data archiving requires attention to the following six interrelated processes:



This handbook reviews each process in turn.

Box 2:

RACcER: Re-Use and Archiving of Complex Community-Based Evaluation Research

The RACcER Project was a collaboration between the Irish Qualitative Data Archive (IQDA) and the Tallaght West Childhood Development Initiative (CDI).

IQDA is a central access point for qualitative (non-numerical) social science data generated within and about Ireland. It was established as part of the Irish Social Science Platform under the Irish Government's Programme for Research in Third Level Institutions (Cycle 4). IQDA is housed in the National Institute for Regional and Spatial Analysis (NIRSA) at NUI Maynooth.

CDI launched a ten year strategy that aims to enable all children in Tallaght West to reach their full potential, improve their health, learning and safety, and to increase their sense of belonging to their community. The first five years is co-funded by the Office of the Minister for Children and Youth Affairs (OMCYA) under its Prevention and Early Intervention Programme for Children (PEIP) and by Atlantic Philanthropies under its Disadvantaged Children and Youth Programme.

CDI has been working successfully with local partners, including eleven primary schools, to deliver a suite of programmes that meet the identified needs of children and families in Tallaght West. Each programme is being rigorously and independently evaluated using a range of methodologies, including randomised control trials, a quasi experiment and process evaluations.

The goal of RACcER was to explore and implement innovative approaches to meeting the ethical and practical challenges involved in archiving and creating appropriate levels of access to the complex qualitative and contextual data generated by the independent evaluation teams that have been commissioned by CDI, without overlapping or duplicating work already being undertaken.

RACcER was co-funded by the Irish Research Council for the Humanities and Social Sciences (IRCHSS) and CDI.

1. Data Management Plan

It is important, both in terms of good research practice and to facilitate future archiving, to ensure that a comprehensive data management plan is put in place at the **onset** of any research project and continued through to data archiving and dissemination.[4] To enable future archiving, best practice is as follows:

[4] For more information on data management models, see the ESRC RELU Data Support Service (http://relu.data-archive.ac.uk/introduction.asp) and the Finnish Social Science Data Archive's pages on data management planning for qualitative data (http://www.fsd.uta.fi/english/data_management_planning/index.html

For the Researcher (Data Depositor):

- Plan to gain and record consent from research participants to ensure that ethical and legal procedures are respected. This consent should include consent to participate in the project and consent to archive the data. Sample consent forms, and a discussion of other issues to be considered, are available on the IQDA website (www.iqda.ie).
- Ensure data are collected in a format that enables archiving and usage (see Appendix 2 for more information). Remember, a key goal of archiving is to facilitate re-use in the **long term** – many file formats used in the past are no longer accessible. In general, open source formats are more accessible than proprietary file formats linked to a particular company or product. If using a digital recording device, ensure it records using WAV files.
- Develop an anonymisation plan to preserve respondent privacy and confidentiality if required under the consent agreement. In most cases, professional ethical standards recommend anonymisation of personal details to preserve respondent confidentiality.
- As the project progresses, identify which text within a document may be particularly sensitive and develop an archiving plan for this data. One approach to developing such a plan is to consider the levels of risk attached to the text across two parameters:

    - ✓ Is there a risk that the text will allow the research participant to be identified?

    - ✓ Is there a risk that the text will be harmful to the research participant? For example, might the text expose the participant to ridicule or have other adverse consequences for them?

By examining these two parameters it is possible to make an overall assessment of the risk level of the data (see Table 1).

Table 1: Assessing Sensitivity Level of the Data

Risk of Identification	Risk Of Harm	Sensitivity Level
Little	Low	Low
Some	Low	Medium
Any	High	High

Once the level of risk has been identified, an archiving strategy can be developed that is appropriate to that level of risk. Common approaches

to dealing with sensitive data include removal of sensitive text segments or placing a time embargo on dissemination of the archived document. Where text has been excised, it is essential that this is noted either in the archived documents, or in the user guide. An example of such a strategy is shown in Table 2.

Table 2: Developing an Archiving Strategy Appropriate to the Level of Sensitivity

Sensitivity Level (See Table 1)	Participant Consent	Anon-ymised	Access	Remove Sensitive Text	Embargo (Anonymised Sensitive Text Removed)	30 Year Embargo (Sensitive Text Not Removed)
Low	Yes	Yes	Restricted	N/A	No	No
Medium	Yes	Yes	Restricted	No	No	No
High	Yes	Yes	Restricted	Yes	No	No

Implementation of the strategy will ensure that the sensitivity levels of the archived documents are reduced following data preparation and processing.

- Create a robust back up procedure to protect data from the threat of technical failure. “321 backup” is a handy rule of thumb: have three copies of your data, on two different types of media, with one version at an offsite location.

For the Data Curator

- Develop clear and easily used data encryption procedures to collect the data from the depositors in a manner that ensures preservation of respondent privacy and confidentiality. We provide information on software for encryption in Section 2, below.
- Confirm with the depositors that there are no legal barriers to archiving their data (e.g. ascertain where ownership of the data lies).
- Confirm that there are no ethical obstacles to archiving (e.g. ascertain the level of consent obtained).
- Identify clearly which person among the depositors (i.e. within the organisation or research team) has responsibility for making decisions about archiving, including access to the data. It is important to remember that archives are created for **long-term** use, and will outlive depositors, whether they are individuals or organisations. For this reason, access agreements need to transfer decisions on access from the depositor to the archive at a specified date. The agreement with the depositor and

the archive must allow for situations in which the depositor ceases to be in a position to make decisions on access.

- Ensure that there are established, clearly understood, and transparent licensing arrangements with the depositors. These arrangements will normally state that (a) ownership of the data lies with the depositor and (b) the depositor confers on the archive the right to disseminate the data for other researchers to re-use. The licence should also ensure that users of the data acknowledge the original producers of the data using a standard citation procedure (for an example of the IQDA depositor licence see Appendix 2).

2. Software Tools

A number of tools exist that can assist both the researcher and the data curator during the archiving process. Those mentioned below have been used, either by the RACcER Project, the IQDA or ESDS Qualidata in the UK.

- **Encryption software**, which is used when receiving and storing the data to ensure that the data are not accessible to unauthorised users. Common encryption programs used in archiving projects include TruCrypt[5], GnuPG or PGP.

- **Automated anonymisation tools** enable a more standardised and rigorous anonymisation process for textual materials. They are useful in ensuring that consistent anonymisation decisions are made, particularly when the data are processed by a team. They do not, however, completely automate the process, and responsibility for ensuring that the anonymisation procedure protects research informants remains with the depositor. All documents anonymised using an automated tool should be manually reviewed before deposit. One such tool has been created by the IQDA and can be downloaded from the website. A similar tool, QualAnon, has been created by the Inter-University Consortium for Political and Social Research (ICPSR) and is available on their webpage (see Appendix 3)[6].

  These anonymisation tools produce two outputs: one fully anonymised document and a second document which contains both the original text and the anonymised text. This is known as a comparison file. The comparison file is useful if the anonymisation decisions are being reviewed by a third party, for example when data processing is being carried out by a research team.

[5] CDI used this software to encrypt their data
[6] http://www.icpsr.umich.edu/icpsrweb/ICPSR/

3. Data Preparation

This phase of the archiving process reduces the levels of sensitivity attached to the data and ensures re-usability. Here standard decisions are made with respect to anonymisation, which can then be applied across the data set. This is a time-intensive process that, like the Data Processing stage, is generally carried out by the researcher or the research team. As such it is wise to build the costs associated with this process into any application for funding. The IQDA will be able to advise on estimating costs.

The key to success is to arrive at an appropriate level of anonymisation that will not reduce the re-use value of the data. The general approach is as follows:

- Remove major identifying data (real names, place and company names);
- Remove all identifying details (names, street-names, real names, occupational details);
- Replace with descriptions that reflect the significance of the original text within the context of the transcript.

The IQDA has produced a more detailed guide to anonymising data, which can be downloaded from its webpage (www.iqda.ie).

Whether or not an automated tool is used, it is best practice to:

- Create a 'Changes File' in which the pseudonyms that have been used to anonymise data are recorded (Table 3). It is particularly important to note any areas where there is doubt as to the best way to proceed. These alerts can be reviewed at the end of the process, in consultation with other members of the research team (if applicable) or with the data curator.

- Use a clear and consistent unique identifier to indicate text that has been anonymised in the documents. This shows where anonymisation has occurred, in order to inform secondary researchers and to facilitate checking and proofing the process of anonymisation itself.

  We recommend the following approach[7]:

  - At the start of the text to be anonymised, use the characters @@.
  - At the end of the text, use the characters ##.

[7] These annotations were recommended to us by Timescapes, an ESRC Qualitative Longitudinal Study

These characters are unlikely to appear for any other reason in the text. For example: “My mother’s name was Mary Brown” becomes “My mother’s name was @@Ann Smith##.”

- Check the data for any statements that may be sensitive and require specific decisions to be made in order to reduce the level of sensitivity (see Table 2). Note these in the Changes File.

Table 3: Example of a ‘Changes File’

Name in Transcript	Nature of Individual or Place (e.g. relationship to respondent, place of work)	Anonymised Name	Names Not Anonymised	Rationale, Special Alerts and Queries
John	Husband	Peter		
London	Husband’s birthplace		London	Place too large for disclosure risk
				“I run a stall selling military stamps on Sundays” Identification risk?

4. Data Processing

Once the data preparation phase has been completed, the next stage is to make various changes, additions, and deletions to the dataset which are required to ensure that commitments to confidentiality are met. Best practice is to:

- Review the sensitive statements noted in the Changes File created in the preparation stage. Alter these statements if possible. If the statement cannot be anonymised, the entire text may need to be removed and explicitly marked as such e.g., @@removed##
- Data require additional contextualizing in order to facilitate re-use. In archiving language this is referred to as ‘metadata.’ Include a header within each document (transcription etc.) giving brief details of that data unit (for example gender of those present, location of interview).

The decision on what to include in the header will also be informed by the confidentiality agreements attached to the data.

- Review the anonymisation decisions noted in the Changes File created in the Data Preparation stage. Anonymise the data following these decisions, either manually or using an automatic tool such as those referred to above.
- Prepare a "User Guide" which gives an overview of the data to assist re-use in the future. This may include the original consent forms, and grant applications. A guide to other material that may be included is available on the IQDA website (www.iqda.com).

If using an automated anonymisation tool:

- Create a comparison document for data verification;
- Note that documents may be incorrectly anonymised or lost (for example, if you automatically replace the name TED with the name SEAN, words that contain "ted" will also be altered, such that "alerted" would become "alersean"). Make a thorough visual check of the anonymised documents to ensure that they have been sufficiently and appropriately anonymised.

5. Data Deposit

Once the data has been processed it is ready to be deposited in the archive. Best practice is as follows:

For the Researcher:

- Deposit final anonymised document as a read-only file. This ensures that the archival document will not be altered by the users, but will allow users to copy and paste as required in the course of their research.
- If it has been necessary to remove sensitive segments from the data set, create a dataset that contains all the sensitive segments. If possible, apply a long-term embargo on access to this dataset. This ensures that in the future researchers will have access to the full data-set, as the level of sensitivity is reduced by the passage of time.

For the Data Curator

- Ensure that research data are stored in a secure environment, with ability to control access.
- Monitor file formats within the archive and ensure that long term preservation plans are in place.

6. Dissemination of Data

Once the dataset has been deposited in the archive, the data curator must manage access to it and promote its re-use.

Best practice is to:

- Have validation procedures in place before granting access to end-users. This will include establishing that users have a comprehensive data security arrangement in place before providing access. An example of an end-user licence can be found on the IQDA website (www.iqda.ie)
- Have in place a standard citation protocol to acknowledge data creation and ownership
- Promote the dissemination of the archived data
- Plan to manage user queries and access on behalf of the data depositor

Appendix 1: Sample Depositors License

Depositor Agreement

Dataset Title:

Conditions of archiving and re-use:
With this agreement and completed online catalogue entry, I/We submit the above mentioned dataset with its supplementary documents to be archived at the Irish Qualitative Data Archive (IQDA). Unless otherwise stated in this agreement, the depositor or the depositing body will retain the ownership and copyright to the dataset and related material. The depositor affirms that the dataset has been prepared in such manner as to protect the confidentiality of individuals and bodies whose details appear in the data, in compliance with professional norms. The depositor undertakes that the dataset does not contravene any laws including but not limited to those relating to defamation or obscenity. IQDA may revise and validitate the deposited dataset, and distribute it for research or teaching purposes under the conditions agreed upon below. IQDA may promote and advertise the dataset in any publicity. Where appropriate, and with the agreement of the depositor, IQDA may make named data (itemized in Condition II, below) openly available for publicity purposes. IQDA takes responsibility for digital storage of the dataset in accordance with data protection and security norms. The archive will ensure that, with the exception of data identified for publicity purposes, the data are distributed for re-use only to persons who have signed appropriate access agreements, and that the conditions set out in this agreement are complied with.

Conditions stipulated by the depositor
I Distribution of the dataset for re-use: access provisions) (Please tick your preferred option)
☐ 1. IQDA may distribute the dataset for *research and teaching purposes* without asking permission from the depositor ☐ 2. IQDA may distribute the dataset for *research and teaching purposes* without asking permission from the depositor, after a specified period of

time agreed in writing with the Director of the Archive. ☐ 3. IQDA may **not** distribute the dataset without the permission of the depositor. (This option should be selected only in exceptional circumstances, after consultation with the Archive staff).
II. **Other conditions or wishes stipulated by the depositor concerning the archiving and re-use of the dataset:**

Name (Block Capitals)	**Signed**	**Date**
Depositor:		_____/_____/_____
IQDA Representative:		_____/_____/_____

Appendix 2: Archiving Formats

Text Formats:

Preferred

* Rich Text Format (.rtf)
* Plain text data, ASCII (.txt)
* eXtensible Markup Language (XML) marked-up text according to an appropriate Document Type Definition (DTD) or schema

Accepted

* Hypertext Markup Language (HTML)
* widely-used proprietary formats e.g. Microsoft Word (.doc/.docx)
* Proprietary/software-specific formats such as NUD*IST, NVivo and ATLAS.ti

Audio Formats:

Preferred:

* Free Lossless Audio Codec (FLAC) (.flac)
* WAV file (.wav)

Accepted:

* MPEG-1 Audio Layer 3 (MP3)
*Audio Interchange File Format (.aiff)

Note. Many commercial digital diction recorders produced by Olympus, Philips and Grundig record in DSS format, which is not ideal if the data are to be archived. When buying a recorder ensure it records in WAV format and at a minimum 44.1k sampling rate.

Appendix 3: Sources of Further Information

ESDS Qualidata
Economic and Social Data Service Qualidata
http://www.esds.ac.uk/qualidata/

FSD – Finnish Social Science Data Archive
http://www.fsd.uta.fi/english/index.html

ICPSR
Inter-University Consortium for Political and Social Research
http://www.icpsr.umich.edu/icpsrweb/ICPSR/

IQDA
Irish Qualitative Data Archive
http://www.iqda.ie

RACcER
Re-use and Archiving of Complex Community-Based Evaluation Research
http://www.iqda.ie

The Respect Project
The RESPECT project was funded by the European Commission' Information Society Technologies (IST) Programme, to draw up professional and ethical guidelines for the conduct of socio-economic research.
http://www.respectproject.org/

The RELU Data Support Service
The Rural Economy and Land Use Programme Data Support Service is a dedicated support service that provides information and guidance to researchers and project managers from the programme on data management, data sharing and preservation. RELU is the data management model used by the ESRC in the United Kingdom.
http://relu.data-archive.ac.uk/

SAI
Sociological Association of Ireland
http://www.sociology.ie

Tallaght West Childhood Development Initiative (CDI)
http://connect.southdublin.ie/cdi/index.php

Timescapes
An ESRC Qualitative Longitudinal Study
http://www.timescapes.leeds.ac.uk