ROER4D Data Publication Guidelines

By Michelle Willmers
(Updated 19 January 2016)

Suggested citation: Willmers, M. (2015). ROER4D Data Publication Guidelines. Retrieved from: http://tinyurl.com/ROER4DDataPublicationGuide

Contents

Introduction: The ROER4D Open Data Initiative
ROER4D contractual framework and data security mechanisms
Articulating a data management plan (DMP)
How to prepare your dataset for publication
- Step 1: Conceptualise the dataset you wish to publish
- Step 2: Identify points of sensitivity in the data
- Step 3: Define an appropriate de-identification approach

Conclusion
Acknowledgements
Useful resources
Appendix A: ROER4D Sub-project Data Management Plan (DMP) overview

Introduction: The ROER4D Open Data Initiative

A principal objective of the ROER4D project is to build an empirical knowledge base on the use and impact of OER from a developing country perspective. The project and the individual sub-projects have an opportunity to demonstrate and expansive version of 'openness' in research by engaging with sharing data sets – anonymised interview and focus-group transcripts, cleaned survey data, and other data sources – openly with the global research community. In line with this ambition, the project has launched the ROER4D Open Data Initiative in which it will work with and assist any researchers wishing to publish data arising from their ROER4D research activity.

Addressing data publication means addressing research data management (RDM) and drawing attention to responsible conduct of research. Funders and institutional authorities are

This work is licensed under a Creative Commons Attribution 4.0 International (CC BY 4.0) licence.
It was carried out with the aid of a grant from the International Development Research Centre, Ottawa, Canada.

increasingly calling for articulation of data management plans (DMPs) – ideally drafted at grant proposal stage – which require researchers to explain how data will be collected and analysed, and to describe ethical or contractual provisions pertaining to the data. DMPs also typically address issues such as intellectual property rights, storage and backup. The decision to publicly release data is only one component of an holistic approach towards data management and responsible research conduct.

Many researchers are concerned about publishing their research data openly, partly because of fears around security and anonymisation but also because of concerns that third-party researchers will misrepresent data or hijack publishing opportunities that may arise from the data. There are also concerns around the time commitment involved in the professional management and publication of data.

Open data publication requiring a significant level of expertise in data preparation and de-identification/anonymisation techniques. In order to address researcher concerns around privacy and ethical issues and ensure high-level technical support, ROER4D has established a formal partnership with DataFirst, a professional data-publishing service based at the University of Cape Town (UCT), in order to curate and publish a collection of ROER4D datasets. DataFirst is part of a small group of elite data services around the world that have acquired the international Data Seal of Approval. This certificate, developed by Data Archiving and Networked Services (DANS) indicates that the platform is trusted digital repository. DataFirst is the only data service in the developing world to have achieved this certification. It is committed to Open Access principles and not only serves as an internationally endorsed data publication service, but also ensures long-term sustainability in hosting the collection (without cost) due to its UCT affiliation.

ROER4D's partnership with DataFirst not only ensures that security and de-identification protocols are strictly adhered to, but also provides access to crucial services such as discoverability (driven by high-level metadata generation), citation tracking, and reportage on usage statistics. The DataFirst Data Portal requires users to go through a once-off registration process, which enables close monitoring of who is accessing the data (and, in some cases, allows researchers to identify for what purpose the data is being used).

ROER4D's Curation & Dissemination (C&D) Strategy and Open Data Initiative have been formulated in line with the IDRC's recently launched Open Access Policy. The IDRC encourages researchers to openly share research data. While this area of activity is not currently required by the Open Access Policy, ROER4D is participating in an IDRC pilot project to develop specific policies and guidelines on open access to research data.

ROER4D contractual framework and data security mechanisms

There are a number of contractual provisions within the ROER4D grant conditions which need to be considered in data publication activity.

The **Memorandum of Grant Conditions** between the IDRC and UCT (the host organisation of the ROER4D Adoption Studies component) as well as between the IDRC and Wawasan Open University (WOU) (the host organisation of the Impact Studies component) provides the foundational set of grant conditions to which all sub-projects (SPs) are accountable. It states:

> *A4.1.(c) The identity of individuals from whom information is obtained in the course of this Project shall be kept strictly confidential.* ***At the conclusion of the Project, any information that reveals the identity of individuals who were Subjects of Research shall be destroyed unless the individual concerned has consented otherwise in writing.*** *No information revealing the identity of any individual shall be included in the final report or in any communication prepared in the course, or as a result, of this Project, unless the individual concerned has consented in writing to its inclusion beforehand.*

This clause should be considered in conjunction with any specific stipulations around disclosure of research subjects that arise in your SP ethics approval process. **Attachment A: Additional Terms and Conditions of the Grant** states:

> *A4.1 Information Gathering*
>
> *Before an individual becomes a Subject of Research, he/she shall be notified of:*
> *the aims, methods, anticipated benefits and potential hazards of the research;*
> *his/her right to abstain from participation in the research and his/her right to terminate the confidential nature of his/her replies.*
>
> *No individual shall become a Subject of Research unless he/she is given the notice referred to in the preceding paragraph and provides freely given consent that he/she agrees to participate.*

With regards to ethical standards, it goes on to state that:

> *In addition to the requirements of paragraphs A4.1(a) – A4.1(d) being complied with, the SP shall:*
>
> *submit the research protocol for the Sub-grant to an appropriately constituted ethics review committee in its institution or at the national level (in the country where the Research Work will be carried out).*

In line with the basic principles of responsible research conduct, the onus is on Lead Researchers to ensure that these protocols are adhered to from the outset of the research process.

If the requisite ethics and consent processes have been adhered to, the C&D team, along with DataFirst experts, will assist interested SPs to prepare their datasets for publication and undertake the de-identification process, but SP researchers will need to ensure that the research subject cannot be identified prior to submission. In order to ensure accordance with contractual and ethics stipulations, the C&D team requires that copies of ethics approval documentation and consent forms be submitted along with datasets for publication. These documents will not be published as a component of the SP dataset, but will be retained for internal record-keeping and verification purposes.

The Impact Studies **2015 Addendum to the Subcontract Agreement** between UCT and SP host organisations also carries a clause which has relevance to data publication. It states:

> *In addition to clause 13.1 [of the Main Agreement], the Parties acknowledge that the sharing of data and the dissemination of the research results to advance the state of knowledge in the field is in the common interest of both Parties. Accordingly, the Parties agree to share data and co-operate with each other and to publish jointly when appropriate. Authorship will be based on commonly accepted academic standards. Where possible to publish in Open Access journals and/or to make reports under a Creative Commons Attribution 4.0 International as the default licence.*

This sentiment is mirrored in the Adoption Studies **Attachment A: Additional Terms and Conditions of the Grant**, which contains a special clause (Section A32) on “Open Content”, stating:

> *So as to ensure the greatest possible development impact, the SP shall ensure that all of the Sub-grant outputs and research findings (excluding computer software) produced during the course of this Memorandum in pursuit of the Sub-grant Objectives are made available to the public pursuant to the Creative Commons Attribution 4.0 License.*

These licensing clauses gives expression to ROER4D’s foundational desire around data sharing, both in and beyond the project, and dictates the CC BY licensing provision which will be adopted in all text and data outputs. The only exception in the licensing context is SP3, which publishes under a CC Attribution-ShareAlike 4.0 International licence (as dictated by their host organisation, Commonwealth of Learning).

In terms of data publication and security mechanisms, the ROER4D Network Hub recognises that data arising from certain SPs may not be suitable for publication in terms of subject-matter sensitivity, even after de-identification has taken place. Note also that there is the option of choosing an embargo period, should any of the SPs wish to publish data before the project end date and maintain exploitation rights for a limited period.

The ROER4D Network Hub provides a long-term archiving service to ensure that all ROER4D reports, outputs, de-identified data, and any other documentation (both public and private) is securely stored for a period of at least five years after the project end date. SPs are also

encouraged to engage support structures in their home institutions around ancillary backup and archiving services (both during and at the end of their projects), noting the stipulation that datasets containing personal identity information will need to be destroyed at project end date unless research subjects have granted permission to the contrary in writing. You can read more about the Network Hub archiving service in the ROER4D Project Archive Overview.

Articulating a data management plan (DMP)

DMPs are articulated by the C&D team for all SPs participating in the ROER4D Open Data Initiative. DMPs are ideally drawn up at grant proposal stage – in part because they reflect aspects of foundational research design; but also because they identify the resources involved for responsible data management and draw attention to any ethical or copyright areas that might be a concern. This component was not part of the research design process in this project, and is therefore undertaken as a key planning component related to publication activity.

The DMP in describes each SP's approach to the following key issues:

- Data collection
- Documentation and metadata
- Ethical and legal considerations (including IP and copyright issues)
- Storage and backup
- Selection and preservation
- Data Sharing
- Responsibilities and resources

ROER4D utilises the Digital Curation Centre's (DCC) DMP tool in order to capture the relevant detail pertaining to the data-sharing activity of each SP. This is done in consultation with SP researchers, and while it is specific to each research project, can be used as a template for future data management planning. DMP statements are useful not only as an administrative mechanism to capture process detail, but also serve as a valuable mechanism for surfacing any challenges or tensions that might exist in the data-sharing process, particularly as relates to resource allocation.

Appendix A provides an overview of the questions researchers will be required to address in order for the C&D team to articulate a SP-specific DMP, along with useful prompting questions and guidance tips.

In addition to SP-specific DMPs, the C&D team has also articulated a "master" ROER4D DMP which outlines the high-level project approach around data management, the detail of which pertains to all SPs.

How to prepare your dataset for publication

The term “dataset” refers to data files as well as any supporting documentation and instruments that comprise the published collection. If you would like other people to be able to use your data, you need to provide as much contextual detail and supporting documentation as possible. ROER4D recommends the following as a minimum:

- **Data files:** Note that when working in Excel files multiple sheets within a document cannot be processed. Qualitative and quantitative files to be prepared separately.
- **Instruments:** Attach research instruments corresponding with submitted data files.
- **Dataset description:** Provide a text-based narrative about the key points relating to your data, highlighting the methods and procedures used to generate the data. This will enable third-party users to establish a context for your data, and assist the DataFirst team in metadata generation.

SP researchers wish to engage in data publication activity should consider the following three steps:

Step 1: Ensure that the requisite ethics and consent components are in place

Conduct a quick audit to ensure that all aspects of the data you wish to share are covered by your ethics approval and informed consent processes. You will need to assess ethics approval and consent documents critically to ensure that there is no wording that prevents public data sharing; such as, “Data will not be shared with any third parties”, or “Data will be used for research purposes only”. Given the CC BY provisions under which ROER4D data is published, the data will be openly available to interested parties in research and non-research sectors. There is therefore no discrimination in terms of who may access and reuse the data; nor are there any stipulations around intended purpose in reuse.

The ROER4D Network Hub cannot publish any data which is not covered by the requisite ethics and consent processes and documents.

Step 2: Conceptualise the dataset you wish to publish

Try to establish a clear sense of why you wish to publish your dataset and what value it might hold in other contexts, whether research or otherwise. Ask yourself the following questions:

- What will the published dataset be comprised of (i.e. which data files and supporting instruments would you like to release)?
- Are there any similar datasets currently published as open data that your work could link to or be aggregated with; or will your work constitute a unique contribution to the field?

- Does your dataset have any special characteristics or unique features which either need to be taken into consideration in the data publication process, or could serve as a hook for drawing attention to the dataset?
- What are your principal concerns in releasing the data?
- Do you have the necessary permissions (ethics clearance, consent forms, etc.) to release the data?
- Are there any technical constraints around data transfer (e.g. file size)?

Step 3: Identify points of sensitivity in the data and define a de-identification approach

Researchers often have a set of concerns about releasing their data. In addition to precautions around ensuring that the identity of subjects or respondents cannot be ascertained, you may wish to consider whether there are any particular constraints or controversial aspects to your data release which you need to safeguard against (such as the identification of institutional contexts or identification of geographical location). These points of sensitivity will inform your de-identification strategy.

IDRC contractual provisions as well as international best practice dictate that all possible measures should be undertaken to preserve the anonymity of research subjects, unless they have specifically agreed to the alternative in writing. It is therefore crucial that a de-identification process be administered on data files prior to publication.

There may be aspects of sensitivity around various aspects of the data, other than just the identity of research subjects. The names of institutions or geographical location, for instance, might be problematic to reveal. “De-identification” refers to the process undertaken to ensure that a person or other entity judged as sensitive in the research context cannot be identified in the data release. As such, it is a means of protecting the author against unintended malicious application of the data in which they might be implicated.

The ROER4D C&D team recommends the following two principal de-identification techniques:

- **Omission:** Whenever disclosive information was included in a response, certain elements of that response (such as a word or phrase) may be omitted so as to obscure identity, while preserving as much of the intended meaning and salience of the response as possible. To make the data as seamless and interpretable as possible, we do typically not provide indications within the data signalling where the omissions have taken place.
- **Revision:** In cases where disclosive information was given but could not be omitted without jeopardising the integrity of the response data, the authors may choose to revised words or phrases to ensure anonymity, while retaining the essential meaning. In most cases, these instances entail changing specific information to something less particular (such as replacing a departmental with a faculty-level identifier).

In order to preserve the integrity of the data, it is suggested that authors utilise [brackets] to indicate where intervention mechanisms have been undertaken for de-identification and clarification purposes. The following scenarios may typically occur:

- **Text revised [ ]:** Brackets indicate that a word or series of words have been altered in a way that preserves original meaning as far as possible. For example, if the original response was: “I work in the sociology department”; this was altered to: “I work in the [Faculty of Humanities].”
- **Note to reader [*]:** Skip-logic was employed in the interview schedule; meaning that there were instances where follow-up questions were contingent on original response. This is indicated in the text as: “[* Skipped. See interview skip logic.]”. In cases where no answer was provided, the authors provide an explanation, clearly marked with the “[* …]” prompt.
- **Interviewer intervention [Int:…]:** Occasionally impromptu follow-up questions not on the interview schedule were asked, creating additional data. In order to distinguish the responses of respondents from the interviewers’ verbal contributions, interviewers’ contributions have been placed in brackets, and on a separate line within the same text block.
- **Unclear to transcriber [unclear]:** In cases where the transcriber was unable to interpret components of a response, “[unclear]” was inserted as an indicator in the text.
- **Withheld [withheld for de-identification]:** In cases where answers to certain questions are judged to be too disclosive (particularly when combined ancillary responses) the authors have withheld certain key responses.

Data de-identification is a crucial and complex process. The C&D team will work with SPs to define a de-identification strategy that speaks to the specific needs of each particular project context. Once researchers have undertaken a first layer of de-identification work in consultation with the C&D team, DataFirst, in its capacity as data publisher, will provide an additional layer of security in undertaking a quality check to ensure that data has been appropriately de-identified and provide advice on how to refine the process.

Conclusion

The ROER4D C&D team is committed to working with and supporting SP researchers who wish to embark on a data publishing process. As a landmark project in the Global South, this data has the potential to serve as a valuable baseline for future investigation, boost the citation of ROER4D outputs, and raise the profile of the ROER4D research community.

It is the belief of the project that a professional approach to curation and open data sharing boosts rigour in the research process and speaks to the advancement of responsible research conduct.

Enquiries around the ROER4D Open Data Initative and data publication process should be addressed to the ROER4D C&D Manager <michelle.willmers@uct.ac.za>

Acknowledgements

Thank you to Henry Trotter for his contribution in articulation of the ROER4D data de-identification approach and to Thomas King for editorial input.

Useful resources

The following resources may be useful when considering data publication and undertaking de-identification processes.

Australian National Data Service (n.d.) De-identifying your data
http://ands.org.au/resource/data-deidentification.html

Australian National Data Service (n.d.) Ethics, consent and data sharing
http://ands.org.au/guides/ethics-working-level.html

Elliott R, Fischer CT & Rennie DL (2010) Evolving guidelines for publication of qualitative research studies in psychology and related fields
http://onlinelibrary.wiley.com/doi/10.1348/014466599162782/abstract;jsessionid=CC325633D29B626651F3027F355EABEC.f01t04

European Commission (2015) Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

McAdoo T (2013) How to Cite a Data Set in APA Style
http://blog.apastyle.org/apastyle/2013/12/how-to-cite-a-data-set-in-apa-style.html

Appendix A: ROER4D Sub-project Data Management Plan (DMP) overview

Data Collection	
What data will you collect?	*Questions to consider:* 1. What type, format and volume of data do you intend to publish? 3. Have you reused any existing data in your dataset?
How will the data be collected or created?	*Questions to consider:* 1. What standards or methodologies have you utilised in your data collection process? 2. How will you structure and name your folders and files? 3. How will you handle versioning? 4. What quality assurance processes will you adopt?
Documentation and Metadata	
What documentation and metadata will accompany the data?	*Questions to consider:* 1. What information is needed for the data to be to be read and interpreted in the future (i.e. what types of documentation will accompany the data to help secondary users understand and reuse it)?
Ethics and Legal Considerations	
How will you manage any ethical issues?	*Questions to consider:* 1. Have you gained consent for data preservation and sharing? 2. How will you protect the identity of participants if required (i.e. which de-identification processes will you adopt)? 3. Do you have any special requirements around sensitive data that need to be accommodated in terms of data transfer and publication?
How will you manage IP and copyright issues?	*Questions to consider:* 1. Who owns the data? 2. Are there any third-party or ownership issues that make the application of a CC BY licence to the dataset problematic? 3. Do you require an embargo period on published data to allow additional time for publication?
Storage and Backup	

How will the data be backed up and stored during the research?	*Questions to consider:* 1. Do you have sufficient storage and backup facilities in your working research context? If so, what is the time period associated with the storage and backup process? 2. How will the data be backed up in your home institutional or organisation context (during and after project end date)?
Selection and Preservation	
Which data are of long-term value and should be retained, shared, and/or preserved?	*Questions to consider:* 1. Are there any data components arising from your research work that you are choosing not publish? What is the reason for this? 2. What are the foreseeable uses (research or otherwise) for your published data? 4. How long will the data be retained and preserved?
What is the long-term preservation plan for the dataset?	*Questions to consider:* 1. Will you be sharing your data on any websites or platforms other than the DataFirst Data Portal? Are there any costs associated with this? 2. Have you costed in time and effort to prepare the data for sharing / preservation?
Responsibilities and Resources	
Who will be responsible for data management and publication?	*Questions to consider:* 1. Who is the principal contact in your SP responsible for overseeing the data management and publication process? responsible for implementing the DMP, and ensuring it is reviewed and revised?
What resources will you require to deliver your plan?	*Questions to consider:* 1. Is additional specialist expertise (or training for existing staff) required in order to prepare your date for publication? 2. Do you require hardware or software which is additional or exceptional to existing institutional provision?