Research Ideas and Outcomes 3: e14672
doi: 10.3897/rio.3.e14672

Data Management Plan

Data Management Plan: IDRC Data Sharing Pilot Project

Cameron Neylon ‡

‡ Curtin University, Perth, Australia

Corresponding author: Cameron Neylon (cn@cameronneylon.net)

Reviewable v1

Received: 23 Jun 2017 | Published: 23 Jun 2017

Citation: Neylon C (2017) Data Management Plan: IDRC Data Sharing Pilot Project. Research Ideas and Outcomes 3: e14672. https://doi.org/10.3897/rio.3.e14672

Abstract

This is the Data Management Plan for the project "Exploring the opportunities and challenges of implementing open research strategies within development institutions" the proposal for which was published as https://doi.org/10.3897/rio.2.e8880. The research proposal calls for support for a pilot project to conduct open data pilot case studies with eight (8) IDRC grantees to develop and implement open data management and sharing plans. The results of the case studies will serve to refine guidelines for the implementation of development research funders' open research data policies.

Keywords

data sharing, data management, RDM, ethics, licensing, DMP

Description

The aim of the IDRC Data Sharing Pilot is to refine guidelines for the implementation of development research funders' open research data policies and to inform IDRC on the design and implementation of its Data Management and Sharing policy. The Pilot funded as an IDRC grant, will conduct open data case studies with seven IDRC grantees to

develop and implement open data management and sharing plans. The case studies will examine the scale of legal, ethical and technical challenges that might limit the sharing of data from IDRC projects including issues of:

- Privacy, personally identifiable information and protection of human subjects.
- Protection of intellectual property generated from projects or potential for financial risks for projects or institutions.
- Challenges in the local legal environment, including ownership of data.
- Ethical issues in releasing or sharing of indigenous and community knowledge, and the relationship between project participants and investigators particularly in the context of historical expropriation of resources.
- Local and global issues of capacity and expertise in the management and sharing of data

The Pilot commenced in October 2015 and will finish at the end of 2016. Case studies conducted with the eight pilot projects will run from March to late November. Each pilot project will be assisted in the process of conducting a data audit, a Data Management Plan (DMP) and the implementation of that DMP.

Data Collection

What types of data will you collect, create, link to, acquire and/or record?

The project itself will generate a range of data types as well as examining data generated by the participating projects. The data covered in this plan is that generated specifically by the investigators in the conduct of the pilot. The data generated in the participating projects is described in their separate Data Management Plans. The specific outputs and data sources from the project covered by this plan are:

- Grant Proposal
- Review of funder data sharing implementation
- Data Audits for each participating project
- Data Management Plans for each participating project
- Correspondence throughout the project with advisors, program officers and investigators
- Interviews recorded and transcribed as part of the review
- Interviews recorded and transcribed with project participants
- Survey responses from participants
- Notes and records of the process of the project

What file formats will your data be collected in? Will these formats allow for data re-use, sharing and long-term access to the data?

The main forms of data that will be collected throughout the projects are:

- Spreadsheets: survey results, forms from participants and records
- Documents: Interview prompts and transcripts as well as notes and records, as well as the booklets and participant materials for the workshops
- Audio/video recordings: Recordings of interviews and workshops
- Images: Photos of the workshops
- Email correspondence relating to the project

Spreadsheets will be maintained as Excel or GoogleDoc formats and exported to CSV for data deposition.

Documents will be maintained as Word or GoogleDoc formats and exported to RTF for data deposition, or PDF in those cases where formatting is significant (e.g the workshop booklets).

Audio files are maintained in a range of formats and will be deposited in an open format, to be determined.

Images related to the project may be shared in some cases. Where this is the case they will be deposited as Tiff files.

Email correspondence may be shared although some content will be sensitive. The format and appropriate repository is to be determined.

What conventions and procedures will you use to structure, name and version control your files to help you and others better understand how your data are organized?

Files are organised into folders by phase of the project and specific outputs. Within the folders the files are named with dates and further relevant information (such as name of interviewee or project). In most cases the data files will be fixed and not subject to substantial editing. Where substantial changes are made an effort will be made to keep both versions (labeled by date) rather than use a formal versioning system.

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future?

Data will be kept in standard and open formats so should remain readable for the forseeable future. The project's formal outputs and reports will be used to index and describe the relevant data files as a record of their place in the project and context. In most cases we will not use a formal metadata schema. One exception is for audio files where the available metadata components will be used to identify, date and describe the context in which the recordings were made.

How will you make sure that documentation is created or captured consistently throughout your project?

Each main deliverable for the project (the published review, interim report, case studies and final report) will have associated data components. The formal publication of these narrative reports will be used as the main trigger for ensuring data is organised and catalogued. All files will be stored in a shared Google Drive which ensures the capture of work and data in progress.

If you are using a metadata standard and/or tools to document and describe your data, please list here.

It is not planned to use a formal standard or related tools to document and describe the data.

Storage and Backup

What are the anticipated storage requirements for your project, in terms of storage space (in megabytes, gigabytes, terabytes, etc.) and the length of time you will be storing it?

We anticipate less than one gigabyte of data and documents to be generated by the project. As far as possible data will be deposited in long term availability archives such as Zenodo, the IDRC Digital Library and the Internet Archive. Deposition will occur at the end of the project or when a relevant formal project output is published. The project partners will not commit to the the archiving of data beyond the end of the project.

How and where will your data be stored and backed up during your research project?

Data and documents are stored on a shared Google Drive folder with at least two local replicas on non-colocated local hard drives.

How will the research team and other collaborators access, modify, and contribute data throughout the project?

The research team, relevant members of the IDRC team, and project participants will be granted access through the Google Drive functionality to folders containing data that is of relevance to them.

Preservation

Where will you deposit your data for long-term preservation and access at the end of your research project?

Long term preservation of openly shared data will be through appropriate long term publish repositories including Zenodo, the IDRC Digital Library and the Internet Archive. In most cases more than one archive will be selected, adhering to the LOCKSS principle. For data that must remain private, primarily some specific audio and transcripts, as well as some notes, we will utilise the IDRC lDigital Library or another appropriate dark archive, to be identified, for preservation.

Data will be deposited with a range of repositories as appropriate. Most of the data is audio, documents and spreadsheets. Zenodo is a natural place to deposit the data, as is the IDRC Digital Library. Some audio and video data may also be placed with the Internet Archive.

Indicate how you will ensure your data is preservation ready. Consider preservation-friendly file formats, ensuring file integrity, anonymization and de-identification, inclusion of supporting documentation.

For all data outputs we will convert any proprietary file formats to open formats for preservation. This will include CSV for spreadsheets, plain text for documents, and audio (to be confirmed, but likely Ogg). We will additionally seek to apply best practice in linking these various objects together as packages of related objects OAIS/OAI-ORE/Research Object tools.

Due to the nature of the project anonymization and de-identification of data from the participating projects is not feasible. Therefore where there are privacy implications or a risk of harm it will be necessary to restrict data access.

Sharing and Re-use

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

Given the nature of the project our aim is to share all data and outputs generated by default. The participating projects have their own ethical and planning requirements which places a limit on the project's ability to share, and specific elements of several projects put considerable constraints on the appropriateness of unilateral sharing. This will be managed on a case by case basis and will form the basis of several of the case studies. In all cases any existing commitments by the projects to their participants and study populations will be observed and respected.

The pilot expects to be able to share most audio/video of interviews as well as transcripts, with a few exceptions, survey instruments and informational materials generated as part of the project, Data Inventories and Data Management Plans for the participating projects (with some exceptions), and formal narrative outputs.

Have you considered what type of end-user license to include with your data?

Where data objects can be shared we will apply the cc0 waiver for data outputs and for narrative documents the most recently available version of the CC BY license. Those data outputs that cannot be shared publicly will only be made available under restricted usage terms to approved users. In most cases these users will be restricted to IDRC staff or the relevant project participants.

What steps will be taken to help the research community know that your data exists?

The pilot will be publicised through formal narrative outputs and through less formal online means. The formally published outputs will be registered through appropriate indices and datasets will be archived in locations providing DataCite DOIs. This will provide discoverability through the main discovery indexes supported by the persistent identifier ecosystem.

Responsibilities and Resources

Identify who will be responsible for managing this project's data during and after the project and the major data management tasks for which they will be responsible.

The collection and management of data during the project is the responsibility of the lead investigator, Cameron Neylon. This includes collecting, cataloguing and managing the various data outputs identified in this plan.

On formal publication the responsibility for management will lie with the relevant publisher and/or repository. There is no long term support for data management after the project concludes. However the relatively small scale of the data outputs and their direct connection to formal narrative outputs makes the ongoing management as simple as it can be.

How will responsibilities for managing data activities be handled if substantive changes happen in the personnel overseeing the project's data, including a change of Principal Investigator?

In the case of a change of PI, responsibility will transfer to the contributing investigator, or to the IDRC program officer.

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

The major costs of data management for the project are in time for management, curation and publication. These are included in the costs of the project under "consultancy" as part of the overall work package outlined in the proposal Neylon and Chan (2016). It is difficult to separate out a specific costs for Data Management Implementation from the overall project due to its nature.

Ethics and Legal Compliance

If your research project includes sensitive data, how will you ensure that it is securely managed and accessible only to approved members of the project?

Access to sensitive data, primarily audio, video and the related transcripts is maintained under access control through the Google Drive folders. Access is restricted to the PI, CI and IDRC program officer. At the conclusion of the project these files will be passed to IDRC for future management. The overall risk posed by these data are relatively small. No data is being held on patients or study populations but only on the projects themselves.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?

There are insufficient resources to support secondary uses of sensitive data. The data collected will be described in narrative outputs and future use or access may be authorised by the holders of data that is not made publicly accessible. As far as possible we will release all data outputs through public repositories.

How will you manage legal, ethical, and intellectual property issues?

Legal, ethical and IPR issues are at the centre of the project. They are therefore subject to an ongoing discussion between the sponsor and the investigators. IP rights for the project are held by IDRC.

Grant title

Exploring the opportunities and challenges of implementing open research strategies within development institutions Neylon and Chan (2016).

References

- Neylon C, Chan L (2016) Exploring the opportunities and challenges of implementing open research strategies within development institutions. Research Ideas and Outcomes 2: e8880. https://doi.org/10.3897/rio.2.e8880