CHED FAQs about Research Data Management

Questions generated for CHED RDM Workshop 18 July 2017, and CHED Collaborative Projects Meeting 13 November 2017. Thanks to Kayleigh Lino for the responses.

- What are the 'right answers' for small scale qualitative data, e.g. interview tapes or transcripts?

This depends entirely on the context of the research. There is no right answer and the decision needs to be made per case.

Transcripts are always a safe option though, since the researcher and/or intervewee can edit out parts if necessary (human identities, sensitive or confidential information, etc).

- What are best storage practices?

This is also dependant on the research context and type of data. Storage decisions need to be decided upon within the context of the research, preferably before the project commences (DMPonline assists you with making the best storage decisions for your research data).

The first thing to ascertain is whether you are obliged to store your data within national borders. Some data is protected by privacy laws that prevent data from being moved beyond national borders. This is often the case with government data. If your data cannot be stored outside of the country, you probably shouldn't be considering cloud storage, since most (if not all) current cloud service providers store your data on servers based overseas… please be aware that data uploaded to ZivaHub is currently stored on cloud storage servers based outside of the country (although there are plans to change this).

It is also important to understand that there are 2 different types of storage considerations when beginning a research project:

1. **Your working storage** is the space that you use to store, edit and work with data while you are collecting/creating and analysing it during the research project. This storage requires careful data management decisions to enable efficiency of navigating through, and working with all of your data throughout the project lifecycle. UCT eResearch provides research data storage services and you can read more about it here.

2. **Your final storage** space where you will archive the data once the project has ended. Final storage generally falls into 2 categories;
   - (1) raw data archive - for storing all the raw data that you collected and analysed, so that you (and your collaborators, or potential reusers) have access to it in the future if you need it.
   - (2) dissemination storage - for storing a processed part of your raw data (the data that directly supports a research publication derived from it), intended for disseminating or sharing with others. When considering where to store and archive your finalised dataset, you need to consider how it will be **preserved**, which entails ensuring that the data files do not become corrupted or outdated (subsequently 'unreadable").

There are some general best practices with respect to storage:

- If you decide to store your data on a personal hard drive, you should purchase at least three drives; use one as the master drive (which you carry with you as you go out into the field) and the other 2 as back-up drives, which should remain at 2 separate, safe locations.
- Perform regular, consistent back-ups from the master drive onto the back-up drives (decide whether to do this daily, weekly, fortnightly or monthly, depending on how frequently you collect data)
- Always store your back-up drives at different locations
- **If you make use of storage services at UCT eResearch/ICTS, all these back-up procedures are done for you.**
- If you decide to store your data in 'the cloud' you will most likely still require the use of a portable hard-drive for collecting and working with data in the field (unless your research is always conducted in wifi connected areas)
- Best practice is to upload data directly to your cloud storage (Google drive/Amazon/Dropbox OR a cloud platform like OSF) from site (or upon collection), but if you're working out of internet connection, you should save all data to your portable hard drive and sync it to your cloud storage as soon as you return from the field
- Since cloud service providers back-up your data for you, you don't have to worry about back-up drives, BUT;
- It is best practice to perform regular back-ups/syncing of your cloud storage onto a local computer or hard drive, in case you delete something from your cloud storage by mistake and cannot retrieve it (this happens!)

● How do we establish a meaningful, decolonial dialogue between: a) Protection (IP, patenting, monetisation, grants), and: b) Openness (Open Access, Open Scholarship)?

No quick answer, requires further discussion… please feel free to suggest some starting points here and we'll keenly plan how to take this further, since it is an NB. discussion…

-

● If a legitimate researcher requests access to data (e.g., UCT, student, UCT academic/supervisor or non-UCT, student), what do they need to provide before sharing data.

This depends on the way in which the author/creator/disseminator of the original dataset has **licensed their data**. As a researcher you have the right to stipulate how others may reuse, cite and share your data.

Most data repositories offer the choice of creative commons licenses, which enable researchers to share their data at varying degrees of openness, while always requiring that the author of the data is formally acknowledged (which enables growth in citations, therefore boosting metrics)

If your data is deposited into a repository as "confidential" (which means others can search and find the metadata and therefore see that the data exists, but only gain access to the dataset through directly requesting it from the author), you are given the opportunity to

define the terms by which you will release the data. You can also state what you require from the requester before you will consider releasing the data.

- How to share qualitative data coded in Nvivo in a form that others can re-use it easily.

This is an excellent question and relates to any data analysed with any proprietary software product. We're not yet as familiar as we'd like to be with using the NVivo software and are plan to find out more about how NVivo tracks changes and queries run across datasets.

Perhaps we could plan a workshop to discuss and practically assess the capability of sharing not only the data processed by NVivo, but the code underlying it.

- What if my PhD student (not NRF-funded) has already received faculty approval for the proposal, and stated that all data would be destroyed when no longer in use?

The RDM Policy makes open access and data sharing the default, but acknowledges that not all data can be shared. In this case, if the student has signed an MoU or contract that stipulates that the data will be destroyed, then they have motivation enough not to share their data.

It is, however, very important to think about why the data is being destroyed and consider whether any of that data might be valuable to future research, not only in your field, but across all disciplines. If you think about how you might be able to share your data from the start of a research project, you might find that it's better to avoid contractual obligations to destroy valuable data and contribute to advancing science beyond your own research.

Note that the RDM Policy enforces data management best practices and **all researchers should be managing their data efficiently, regardless of whether they will share it or not**. Data management and data sharing are separate endeavors. All research data should be managed, even if it will eventually be destroyed, so the researcher should still be encouraged to create a Data Management Plan (DMP) and implement data management practices into their research project, to enable efficiency of data collection, analysis and collaboration. The DMP will then describe the motivation for not sharing and/or destroying the data at conclusion of the project.

- What is the benefit for the researcher?

Many benefits!

1. Satisfy funder requirements and get funding

Increasingly, funding bodies mandate the submission of a Data Management Plan (DMP) to ensure that data can be preserved and shared.

A detailed DMP will help you:

· meet the requirements of your funding agency
· secure funding
· address ethics, preservation, documentation and verification issues

2. Satisfy journal publishers' requirements and get published

Increasingly publishers are requesting that you share your data on OA repositories in order for reviewers to check the reliability of the data that informs the research being submitted for publication

3. Organise and understand your data

By managing your data, you make it easier to understand the details and procedures relating to your data and data collection throughout the life cycle of the project.

Good data management makes research easier

4. Collaboration efficiency

Well managed data ensures that all research project stakeholders/collaborators can navigate, understand and use the data efficiently. Data management also enables the fluid movement of stakeholders in and out of a project - new collaborators can pick-up on previous project members' work easily and without confusion.... time saving!

5. Increase citations and get recognition

The data you collect are the basis of your research. By managing your data you increase your chances of being recognised and cited by others.

By sharing your data you're essentially publishing an additional research output that can get you additional citations and increase your metrics.

Well-managed data ensures that your research can be shared, reused and validated by others.

6. Enable reproducibility and growth in research output

Well-managed data is:

- findable
- accessible
- interoperable
- reproducible

If your data is well managed, the time and cost of future research efforts are greatly reduced. By managing and sharing your research data, you reduce duplication of research efforts, enable growth in future research output and facilitate new discoveries.

- What data must a share, for example, “raw” data or cleaned transcripts or analysed transcripts? Can I remove sensitive information? (e.g. in students’ visual stories). Do I share my full dataset, even if I did not use it? Or just that relating to a publication or thesis?

The funder and journal mandates require that you ***share*** *only the data that you used to inform your research publication*. However, ***all*** *data should be* ***managed*** throughout your research project, and if you wish to share all or a part of your raw data, there are services in place to enable you to do so (osf.uct.ac.za).

The RDM Policy makes open access and data sharing the default, but acknowledges that not all data can be shared. You have the right to de-identify and edit out parts of the data that you deem unethical to share. It is important to keep track of the changes that you've

made to the data and to explain what has been changed (and why) in a document accompanying the data, as a part of the upload/deposit, or in the mandatory description field.

- Qualitative data takes on meaning within a particular context - how does one communicate this context along with the data?

I think this question was discussed quite well during the workshop, but in summary: You can deposit and share any amount of additional information in any format along with your research data. You could keep journals, diaries, AV recordings of the context and interactions carried out during data collection and deposit these along with the data. You can even write philosophical reports describing the context and experiences of the data collection process. In short, the more context you can provide, the better.

- How do I ensure that my data is referenced appropriately? What are the copyright issues?

Almost all data sharing platforms (OA repositories) enforce the addition of a license to the data upon upload, so that the depositor chooses a standardised license (read more here) that states how the data may be reused and shared.

Once the dataset appears on the website, it always has citation options that state exactly how the data should be cited when reused or referenced, so that the original creator gets academic credit for their work.

Most OA data sharing repositories assign a DOI (or persistent ID/handle) to your dataset, so that wherever it is used and referenced, it is always trackable and traceable back to the original.

Lastly (and this may come as a shock), raw data is not protected by copyright law in South Africa. This is extremely problematic and requires much more research, but essentially, if you take all the precautions of licensing your data appropriately, specifying citation and acknowledgment terms, and adding additional contextual descriptions, if somebody uses your data against your will, there is no legal way to punish them. BUT it is unlikely that such offenses will come into play and current evidence shows that the positive reuses of data are vastly greater than the negative.

- How do I control how my data is used and prevent abuse?

Essentially, you can't control how people reuse your data, but you can put measures in place to prevent abuse. Licensing and public acknowledgements that get published as metadata accompanying your data allow you to publicly state how the data should be interpreted and used.

Remember that you don't have control over how others reuse and interpret your research publications either. Many citations in research are taken out of context, and in other cases are used to support a contradictory argument. Like with publications, research practice invites us to engage in a dialogue and comment upon other interpretations of our research.

If you have licensed your work and provided appropriate description that situates your data within the context that you intend it to be read, then you have ethically disseminated your work and any abuse of it can be traced back to your original intentions.

- Might scare potential research participants off if the consent form indicates that data will be shared?

This is a very real concern and will be approached differently depending on the type of research being conducted and the kind of subjects involved. Consider the other side of the coin, where data sharing might provide a research participant with the opportunity to be afforded a public voice where they would otherwise be censored or silenced... many research participants won't be scared off, while many others will refuse to participate. It really depends on the situation and how you choose to communicate your research intentions to your participants. There are ways of working around the issue and discussing the pros and cons with your research participants, but if data sharing is definitely going to affect the results of your research negatively, you have motivation not to share.

- What about research participants who do want to be identified in the research?

Partly answered above in " Might scare potential research participants off if the consent form indicates that data will be shared?" If your research participants want to be identified this is strong motivation for you to ensure that your data is efficiently managed so that you can share your data.

- How will researchers deal with the added load?

It might take some time in the beginning to wrap your head around all the new info and get used to the many tools available to assist with data sharing and management, but RDM practices have proven to save researchers and research groups a lot of time in the long run. Essentially, if you manage your data right from the start, it is much much easier to navigate, work with, analyse, and if you want to, share it.

At the very least, filling in a DMP template might seem a lengthy process in the beginning, but it will definitely save you a lot of time and extra effort during your research project, because you will have identified all the risks and made plans to mitigate them before they occur.

EXTRA WORK?

· RDM makes it easier to understand, preserve and work with your data
· RDM assists you with planning your research project in order to satisfy funder requirements.
· RDM validates your research and eases the transition into a research project for new members or collaborators.
· Data transformation can lead to unforeseen uses in other research disciplines: new use cases = more citations and greater reach / recognition of your work… you never know how valuable your data might be to another researcher in another discipline!

- Where do I deposit my data?

ZivaHub: Open Data UCT is the new institutional Figshare repository for managing and sharing data. There are also several other options, and it is up to you to decide where you'd prefer to deposit and/or share your data: See UCT DLS data sharing guidelines here.

- What form should it be deposited in?

We've discussed the topics of what type of data you should share and how you can describe it in FAQs above.

You should always try to convert proprietary file types (files produced from paid-for software) into open formats (txt, xml, csv, pdf, jpg, mp3, mpeg4) and deposit the open formats with your proprietary files.

You should always describe as much as possible about your data: where and when it was produced/collected, what software or instruments you used to create/collect it, what tools are required for processing the data, etc… Imagine you're explaining how you collected, processed and analysed your data to somebody who knows nothing about your research or discipline.

See our slideshow on RDM services, tips and best practices here.

- Does the draft UCT policy mean that I have to make my data available?

The UCT RDM Policy requires that UCT researchers *manage* their data efficiently according to best practices. The RDM Policy makes open access and data sharing the default, but **acknowledges that not all data can be shared**. If you have strong motivation not to share your data (perhaps your research contract states that you will not share it), you are not entitled to do so.

ZivaHub, UCT's new institutional data repository, allows researchers to "share data" in several *restricted* ways:

- You can create a metadata record (description of your data) and publish just the metadata without depositing any data at all
- You can create a metadata record and deposit the data as "confidential" so that only you can see the data, but you can generate a private link to enable access to specific people
- You can upload your data with metadata, but choose not to publish it, so that you have access to your data archive, but nobody else does

Additional questions from CHED project meeting (13/11/2017)

- Purpose of the plan?
  To assist you with efficiently planning how to collect and manage your project's data, in order to make it easier for you and you research project's stakeholders/collaborators to work with and analyse it further down the line

- How to answer questions like storage for sensitive data
  No correct answer, you choose how you store your data and the DMP template guides you to make that choice

- Who's the plan for?
  Primarily yourself and your present/future research project collaborators

- Data lifecycle

- what is the recommended lifetime?

Anything between 5 years and into perpetuity. Some universities have a 5-10 year policy, but UCT hasn't stipulated a lifespan yet

- scale vs. volume (requires more explanation on DMPonline)

These are words generally used when referring to "big data". They represent two different types of "big":

1. lots of files, different types of data, many folders = big data with a complex structure (requires efficient data management protocols)
2. One folder, one type of data, massive files (TBs) = big data with simple structure, but complex HPC analysis requirements (requires efficient storage protocols)

- Levels of access

  There are different ways in which you are still able to share parts of sensitive datasets:
  - You can create a metadata record (description of your data) and publish just the metadata without depositing any data at all
  - You can create a metadata record and deposit the data as "confidential" so that only you can see the data, but you can generate a private link to enable access to specific people
  - You can upload your data with metadata, but choose not to publish it, so that you have access to your data archive, but nobody else does

- Difference between outline & full DMP

- Ethics alignment

  We are currently in the process of engaging with the Office of Research Integrity, in order to integrate data management and sharing practices into the process of ethics clearance process

- What is the researcher's responsibility wrt monitoring reuse?

  The RDM Policy does not stipulate that researchers are required to monitor the reuse of shared data - you can never control what others might do with your data, or any other research output. The researcher is responsible for ensuring that they have the right to share their data, stipulating how the data can be reused, and making sure that their data is deposited in accordance with the UCT the terms of deposit.