Data Management Plan Exemplar #3: Mixed Methods

Fictional Exemplar

Administrative Details

Project Name: Portage DMP Exemplar - "Mixed Methods"

Project Identifier: 8675309

Grant Title: 12345

Principal Investigator / Researcher: James Doiron

Project Data Contact: James Doiron (jdoiron@ualberta.ca)

Description:

This is a fictional exemplar data management plan focusing on mixed methods research (survey & interviews/focus groups). This exemplar DMP was developed in February 2020 by James Doiron (Co-chair, Portage DMP Expert Group) for educational and guidance purposes. For these purposes, the premise of the project being described is that the study topic of interest is not sensitive in nature, and that participants are consenting for their de-identified information (data) to be deposited for long-term preservation, discoverability, and open access.

Institution: University of Alberta

Data Collection

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

We will be collecting surveys which will then be exported into tabular format.

We will also be conducting both semi-structured interviews and focus groups that will produce both digital audio and text (transcriptions) based data.

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?

Our file formats will exist both in non-proprietary and proprietary formats. The non- proprietary formats will ensure that these data are able to be used by anyone wishing to do so once they are deposited and made openly available.

Surveys will exist in .csv (non-proprietary), MS Excel, & SPSS (both proprietary) formats. For more information regarding SPSS see: SPSS Wikipedia https://en.wikipedia.org/wiki/SPSS

Interviews & focus groups data will exist in .mp3 (non-proprietary), MS Word & NVivo (both proprietary) formats. For more information regarding NVivo see: NVivo Wikipedia https://en.wikipedia.org/wiki/NVivo

Any survey data deposited for sharing and long-term access will be in .csv format so that anyone can use them without requiring proprietary software.

The final de-identified versions of the interviews and focus groups transcripts will be exported into a basic non-proprietary text format for deposit, long-term preservation and access.

If data are collected using laptops or mobile devices, please explain how you will securely store and transfer the data.

Laptops are not being used for any data collection, though encrypted digital voice recorders (DVRs) will be used to collect both interviews and transcripts. Interviews and focus group digital audio files will not be stored on the DVRs, only collected and then securely transferred to the project's cloud based virtual research environment space via a secure FTP (File Transfer Protocol).

Data Management Plan Exemplar #3: Mixed Methods

Fictional Exemplar

Survey data will be collected using the REDCap software platform hosted by the Women and Children's Health Research Institute **(WCHRI)** located at the University of Alberta.

When using acronyms within your DMP it is always recommended to spell them out at least once within any given category so that readers understand what it is.

REDCap is a browser based electronic data capture software that is widely used for academic research purposes and supports the development and implementation of surveys. WCHRI securely hosts a REDCap instance that is made available to researchers located at the University of Alberta. Survey participants will use a secure and unique web-based link to access the survey, and all information they provide will be entered and saved directly into the REDCap system. For more information regarding WCHRI's REDCap instance see: WCHRI REDCap https://www.wchri.org/redcap

Interview and focus group data collected will be securely transferred to the project's virtual research space within 48 hours of being collected using a secure file transfer protocol (SFTP) platform.

It is considered best practice to clearly define the length of time that data will be stored on mobile devices (e.g., laptops, electronic pads, digital voice recorders), as well as how they will be transferred to your research space.

These data will be stored and accessed by project researchers and staff within the project's virtual research space located on Compute Canada's cloud Rapid Access Services (RAS) system. For more information regarding this service provided by Compute Canada see: Compute Canada's cloud Rapid Access Services https://www.computecanada.ca/research-portal/accessing-resources/rapid-access- service/

How much data do you anticipate collecting? Include an estimate of how much storage space you will require (in megabytes, gigabytes, terabytes). This estimate should also take into account storage space required for file versioning, backups, and the growth rate over time.

We estimate that we will be collecting approximately 800 surveys, 20 interviews (approximately 30 min in length each), and 2 focus groups (approximately 90 min in length each). Total magnitude of data, including accounting for versions (raw, master, analytic) is estimated to be under 30GB.

Estimating as early as possible how much data you will be collecting is important as this can help you to have a clear sense of how much storage space you require both during the active phase of your research project, as well as to support the long-term preservation of your data.

Are there any existing data that you can re-use? If so, please explain how you will obtain that data and integrate it into your research project.

We have searched for existing data that may be re-purposed or integrated into our project but were unable to find any. We searched a wide range of disciplinary, institutional (Dataverse) and national repositories (Federated Research Data Repository (FRDR), Stats Canada public use microdata files, etc.). We also contacted the data team located within the University of Alberta Libraries and had an in-person consultation with Anna Bombak who was extremely knowledgeable and helpful!

What conventions and procedures will you use to structure, name and version control your files to ensure that your data is well-organized?

Having clear file naming and versioning protocols will help to ensure that your data are well organized, that participant confidentiality is maintained, and support all research activities, including data collection, data processing, analysis, and deposit!

Protocols and procedures relating to file and folder structure, naming conventions, and version control will be developed and implemented, and all study staff and researchers will undergo training regarding these.

Upon export from REDCap, survey data will be named as 'raw' data. While there may be some additional and yet-to-be-determined versions in between, ultimately the final and processed data will be named and versioned as 'master' data. From the master survey data, a variety of analytic data files will be created, with each adhering to a file naming convention that will be developed to support both management and analysis of data and that will be clearly outlined within the aforementioned protocols.

For our qualitative research data - interviews and focus groups - a similar process will occur. Digital audio files will be named and versioned as 'raw audio', while the verbatim transcripts produced from these will be named and versioned as 'raw transcript'. Once a raw transcript has been processed it will be named and versioned as a 'master transcript'. Copies of the master transcripts will be created and used for analysis purposes.

Documentation and Metadata

What documentation will be needed for the data to be read and interpreted correctly in the future? This includes study-level documentation, data-level description, and any other contextual information required to make the data usable by other researchers

Survey data will be collected within REDCap, a secure electronic data capture and management software hosted by the Women & Children's Health Research Institute (WCHRI) at the University of Alberta. REDCap features include the ability to develop and export a data dictionary which will outline all codes and variables within the survey. Key documentation related variables will be automatically populated within the survey data including time and date stamps and other key information to support data management and analytic activities.

All qualitative interviews will include summary information including data collector, location of interview, and the date that the interview was conducted. Additionally, qualitative interviews will have accompanying field notes containing key contextual information and metadata.

File naming documentation will be developed and implemented. Components of file names will include as needed: file version (raw, edit, master, analytic), date (i.e., dd/mm/yyyy), and any applicable key contextual information, (e.g., geographical location, interviewer initials or code).

Please list the metadata standard and tools you will use to document and describe your data. If there is not an appropriate standard, please explain how you will ensure consistency in your documentation.

After consulting with the metadata team located in the University of Alberta Library, we concluded that the Data Documentation Initiative (DDI) metadata standard would optimally support our survey data, and especially as our data collection platform, REDCap, also employs DDI metadata standards. More information on DDI can be found at: https://ddialliance.org

It is not yet determined what metadata standards will be used for supporting qualitative research. We will be consulting with on-campus expertise within the University of Alberta Libraries, other qualitative researchers, as well as the Qualitative Data Repository (QDR) located at the Maxwell School of Citizenship and Public Affairs at Syracuse University

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

We will develop and implement clear metadata documentation protocols prior to starting data collection which will clearly communicate expectations, standards, and processes for capturing and implementing documentation to support the research project. Example topics that these protocols will cover include file naming conventions, file versioning, folder structure, and both descriptive and structural metadata. Researchers and staff will have the opportunity to contribute to these protocols, and once they are finalized, they will be implemented to support the active phases of the research project. Documentation will be kept in clearly named folders within the research project folder space and will be accessible by all researchers and staff.

Storage and Backup

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

All data storage and back-up procedures will be clearly outlined within the project's data collection policies and procedures which will be developed prior to data collection.

Survey data will be collected using REDCap, a secure data collection and management software hosted by the Women & Children's Health Research Institute (WCHRI) at the University of Alberta. REDCap servers undergo regular (daily, weekly, monthly) backups. Secure transferring of data from the REDCap platform to our virtual cloud-based research environment will occur using both encryption and a secure File Transfer Protocol platform. Upon being exported from REDCap, survey data will be immediately ingested into our research project space located on Compute Canada's cloud platform and which has regularly scheduled backup processes in place.

Qualitative interviews will be conducted using encrypted digital voice recorders. Upon completion of interviews they will be securely transferred within 48 hours to the virtual research project space located on Compute Canada's cloud platform. Once the audio interviews are uploaded to the cloud platform, they will be deleted from the digital voice recorders.

How will you ensure that sensitive data is stored securely and only accessible to the research team during the research project?

All data will be securely stored in our virtual project space located on Compute Canada's cloud platform and these will be accessible only by approved researchers, trainees, and study staff. Access to the platform is securely password protected, with access rights ultimately approved by the Principal Investigator and assigned by Compute Canada.

For more information regarding this service provided by Compute Canada see: https://www.computecanada.ca/research-portal/accessing-resources/rapid-access- service/

Preservation

Which data are selected for preservation and access will depend on potential reuse value, whether there are obligations to either retain or destroy data, and the resources required to properly curate the data and ensure that it remains usable in the future? In some circumstances, it may be feasible to preserve all versions of the data (e.g. raw, processed, analyzed, final), but in others, it may be preferable to only keep only selected data (e.g. transcripts instead of audio interviews).

All data will be maintained for a minimum 5 years after the completion of the project, as per University of Alberta ethics requirements.

For the long term, we will be preserving both the raw and the master (cleaned and processed) versions of the surveys. The de-identified and processed versions of surveys will be deposited for long-term preservation and open access. We will be preserving only the de-identified interview transcripts, and these too will be deposited for long-term preservation and open access.

At the end of your research project, where will you deposit your data for long- term preservation and access?

At this time, we believe that we will be using the University of Alberta's institutional data repository, Dataverse, to deposit and support the long-term preservation, discovery, and access of our data. Dataverse is freely available for use by our project and contains a number of desirable features including the assignment of unique and persistent digital object identifiers (DOIs), the ability to restrict access to data at the file level, built in data citations, data usage metrics, and file versioning. More information regarding Dataverse is available at: https://en.wikipedia.org/wiki/Dataverse

Should the UofA Dataverse be determined to not be able to support any of our data deposit needs, we will be consulting with the UofA Library data team to help guide us through the deposit process in order to find the optimal solutions and support for our project.

Please describe how you will prepare the data for preservation and access, including any necessary procedures for data cleaning, normalization or de- identification. Explain how you will prevent data from being lost while processing and converting files.

The surveys collected are anonymous, with no direct identifiers present. Survey data will undergo data quality checks, including missing and out of range values, logic rule and skip pattern violations, and derivation of scales. When converting survey data between formats there will be systematic checks - both case and variable wise - to ensure that no data are lost.

The audio interviews will be transcribed by research assistants - this will be a verbatim transcription, with no contextual information removed. However, interview participants will be de-identified by referring to them by predetermined codes or pseudonyms. Once the initial verbatim transcription is completed it will be saved and versioned as a 'raw' transcript. This raw transcript will then be provided to the interviewer to review for completeness and to remove any necessary indirect identifiers from the text. Once this has occurred it will be named and versioned as a 'master' transcript.

Data Sharing and Reuse

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final). Consider which data may need to be shared in order to meet institutional or funding requirements, and which data may be restricted because of confidentiality/privacy.

We plan to share processed survey data. This will include responses at the participant level, with all direct identifiers removed, and including derived variables used for analytic purposes.

Regarding the qualitative interviews, we plan to share processed and de-identified transcripts.

We will obtain participant consent for sharing these data and will additionally be depositing and making available an example of both our study information letter and participant consent form, as well as our approved institutional ethics application.

How will you be sharing your data? (e.g. institutional repository, a specialized data archive, project website, informal/on-request sharing). Include a brief description of any resources needed to share your data (equipment, systems, expertise, etc.).

We plan to share our data using the University of Alberta institutional data repository, Dataverse. Dataverse is freely available for use by our project, and contains a number of desirable features such as the assignment of unique and persistent digital object identifiers (DOI's) which directly support data discoverability, citations, and usage metrics; the ability to restrict access to data, file versioning, and customized terms of use.

Should Dataverse be determined to be able to support any of our data deposit needs, we will be consulting with the UAL data team to help guide us through the deposit process in order to find the optimal solutions and support for our project.

Please describe whether there will be any restrictions placed on your data when they are made available and who may access them. If data are not openly available, describe the process for gaining access.

All de-identified data will be openly discoverable and accessible to anyone via the University of Alberta Dataverse repository platform.

What type of end-user license will you include with your data? Please include a copy of this license with your Data Management Plan.

All data will fall under open data licensing (CC-BY)

Responsibilities and Resources

Who will be responsible for data management during the project? (i.e. during collection, processing, analysis, documentation). Identify staff and organizational roles and their responsibilities for carrying out the data management plan (DMP), including time allocations and training requirements.

It is important to think ahead and identify who will be involved in the management of your data. Not all projects are the same but, to varying degrees, all may require dedicated resources to help effectively manage the data!

Research data management will be a shared responsibility and will involve the Principal Investigator, co-investigators, collaborators, trainees, and research staff. That said, we additionally plan to have a dedicated research data manager position, estimating this to be at a 0.6 FTE position. We will be including this position into our funding application and budget.

We will be creating a Research Data Management Committee (RDMC) that will be responsible for data governance, including developing policies and procedures relating to research data management and that will be implemented throughout the project. We plan to conduct a 1-1/2-day research data management training event prior to any data being collected, and all investigators, trainees, and research staff will be required to participate in this training.

All research staff involved in collecting data (both survey and qualitative interviews) will sign a confidentiality agreement. Analysts, investigators, and trainees that will be handling data for analysis and dissemination purposes will additionally undergo training and signing of confidentiality agreements.

What will happen when personnel changes occur or if the principal investigator leaves the institution before the project has concluded?

Should personnel changes occur we will refer to our Research Data Management Committee (RDMC) as well as the RDM Policies & Procedures governance. Together, these will inform us of optimal coverage by identifying both primary and secondary individual roles and

responsibilities for various RDM activities. In the event that new team members/research staff/trainees come on board, we will similarly refer to these policies and procedures in providing direction for supporting optimal transitions. All new members of the research project will undergo in depth RDM training and signing of confidentiality agreements.

Should the PI leave the institution prior to the conclusion of the project, the existing PI will work closely with the RDM Committee, as well as institutional administrative research support, to identify the optimal scenario for supporting the research and management of its data moving forward while adhering to all ethical, legal, and contractual obligations. All RDM policies and procedures, including plans for the deposit and long-term preservation and appropriate access of research data will remain intact.

Who will be responsible for data sharing and preservation after the project has concluded? Indicate the party who will have primary responsibility for how the data will persist over time when the original personnel have moved on.

The responsibility for data sharing and preservation after the project has concluded will be a shared one. Ultimately, the Principal Investigator is responsible for ensuring that these data move forward with being preserved and appropriately shared, but this will occur with the support of the RDM Committee, as well as collaboratively with the University of Alberta Library.

Data will be deposited into the UofA Library institutional research data repository, Dataverse. All data deposited will be open access and, as such, UofA Library will effectively become the stewards of these data, ensuring their continued preservation and access.

What resources will you require to implement your plan? Will extra people, time or hardware, storage be required? How much will this cost (estimation)?

The resource needs that we have currently identified are:

- A dedicated research data manager for a period of 12 months, estimated to be at a 0.6 FTE position. We will be including this position into our funding application and budget at an overall amount of $42,000.

This is simply an example demonstrating where a project has deemed that a dedicated research data manager position is necessary. Not all projects will need, or can justify, having a dedicated position such as this. It is still important though to identify a project's RDM needs and from there determine what supports may be required as well as to identify who among the research team will be responsible for specific activities (e.g., investigators, research staff, graduate students).

- Cloud based virtual research environment space. This will be provided by Compute Canada via their Rapid Access Service (RAS). See Compute Canada Rapid Access Service https://www.computecanada.ca/research-portal/accessing-resources/rapid-access- service/
- Proprietary analytic software both for the survey (SPSS) and qualitative data (NVivo). We are not yet sure if we will need to pay for license costs for these, and so will be consulting with Compute Canada on this. If so, we are estimating that we will require 3 SPSS licenses and 2 NVivo licenses for a period of 1 year and will need to determine the cost of these as we would require server-based licenses. We will determine these costs and include them within our funding application budget.
- We will require three encrypted digital voice recorders. We have identified some makes and models that will meet our needs and are budgeting a total cost of $750 CAD.

This document was generated by *DMP Assistant* (*https://assistant.portagenetwork.ca*).