Group Project
Data Management Plans
INFO 6540 Database Management Systems
Jillian Pulsifer, Amy LeMoine, Cassandra Larose, Kristy Hancock
(JACK Data Consultants)

Data Management Plan Proposal

Project Name: Professor Green

Plan created by JACK Data Consultants

Principal Investigator of the Research: Professor Green

DATA COLLECTION

The project entails collecting data from healthcare organizations, open data, and conducted interviews. The types of data that will be collected and created are text, numeric, and audio data. The file formats that will be collected and created are Microsoft Word .doc, .pdf, Excel .xlsx, and .txt files. The audio files will be recorded in .mp3 format, and open data collected will be in. tsv file format.

The JACK Data Consultants (JDC) team will review material from the binder of printed Excel spreadsheets from 2002 to determine if any data is pertinent to the current research project. Relevant data will be scanned and added to other data; inactive data can be discarded or stored for future projects.

File Naming Conventions

The recommendation is to implement a standard file naming system for all files. This will assist in organizing files, allow organization by date, and maintain version control.
The naming convention may be based upon this format: QEII_Staffing_20140409_v01.docx

- Use YYYYMMDD for the date.
- For each hospital or organization use a unique identifier and be consistent when naming files.
- Keep track of versions using v01, v02, etc.

When using folder systems ensure folder hierarchies are as simple as possible. If the client wishes to use folders, they may select whichever organizational method is preferred: the recommendation is to file by type of file (e.g., Interviews), then by hospital (e.g., IWK Hospital). E.g., Healthcare_Expeditures>QEII>datasets

DOCUMENTATION AND METADATA

Maintain full and complete documentation for the study to ensure that it can be interpreted correctly in the future. The following elements should be included:

- Research topic (e.g., teamwork in hospital environments, specifically interdisciplinary primary care teams)
- Description of method of data collection
- Definitions of variables
- Explanation of data coding practices
- Format and file type of data that is being collected and created (including changes to file formats)
- Details of contributors and responsibilities

Standards for data coding must be created and used throughout data collection and during the research project. Determine short codes for the institutions (Hospitals, e.g., Halifax Infirmary = QEII). Ensure strict confidentiality practices are maintained when describing interview participants (e.g., used coded names for each participants and remove/change identifying features).

Metadata Standards

Ensure that information is consistently captured to guarantee uniformity and completeness. During the research project, use a template when conducting textual analysis from outside sources. JDC can assist with creating this template. (The suggested method is to use a standard Excel sheet for data entry.)

To provide consistent metadata to the Excel spreadsheets, use a tool like **Colectica** (https://www.colectica.com/software/colecticaforexcel/), which allows for metadata documentation input directly into Excel, which imbeds the metadata into each file. It includes the metadata elements: Title, Creator, Description, Data type, Variable descriptions. There is a free version of Colectica, with the option for upgrades for more features. Colectica is exportable to an XML file in the DDI metadata format, the standard for data documentation (http://www.ddialliance.org/training/why-use-ddi). This standard is machine-readable and interoperable, encourages comprehensive data description, and enables the reuse of metadata. DDI is appropriate for social and behavioural science research and uses XML.

STORAGE AND BACKUP

Anticipated Storage Requirements

The current amount of collected data is 24 GB and it is projected to triple over the next 10 years, to a total of approximately 72 GB of data. JDC therefore recommends obtaining 100 GB of total storage to allow for growth over time. This will also allow data to be stored for an indefinite amount of time.

Data Storage and Backup Plans

Standard data practice is to store data in three separate places in the event of natural disaster, file corruption, or human error. Also store all data in multiple formats. Thus, it is recommended that the research data be stored in three locations to avoid the loss of your important research.

1. **Secure Cloud Storage. Box** is a secure cloud storage service. The Box Starter package (2018 version) would facilitate storage and access needs, and costs $7/user/month. Box is the best option for cost, security, and maintaining Canadian privacy. The plan has a minimum of three users but allows the owner to provide granular access to specific persons; the Principal Investigator maintains complete control. This software provides both storage and backup features and offers up to 100 GB of storage.

   For a larger budget, the Box Business package (2018 version) has additional security features such as Advanced user and security reporting and monitoring. Given the sensitivity of interview data, this may suit the researcher better. The cost is $21/user/month and provides unlimited storage and the option to add more collaborators.
   Box provides *Box Zones*, so the client can preserve data inside Canada, preserving Canadian privacy requirements.

2. **Network Drive.** Another copy of the data could be kept on the **Dalhousie Network Drive** that the Principal Investigator has access to as a member of the Dalhousie faculty.

3. **External Hard Drive.** A third copy of the data could be stored on an external hard drive that can be kept in a secure off-site location. Although it is good practice to perform backups after every change made to data, JDC recommends creating a backup schedule and ensuring that files are backed up weekly, or monthly, depending on how often the data is updated. Data should be encrypted on the hard drive to ensure complete control over any data that contains personal or sensitive information.

Data Access, Modification, and Team Communication

Currently, the research team communicates through Zotero (for documents), Dropbox (for audio files), and Google Docs (for transcriptions). Box may be used in lieu of these three programs to streamline team communication. Box offers features such as large file sharing, real-time editing and automatic version control. The password protection feature will also allow the Principal Investigator to share only anonymized data with graduate students, or other specified collaborators. Additionally, team members should refrain from discussing confidential matters or divulging identifying information in Box or any other means of digital communication, including email.

PRESERVATION

Long-term Preservation and Access

Upon completion of data collection and at the end of the research project the final data should be shared with colleagues and other researchers, as it will have enduring value to future researchers. The information can be entered into the Dalhousie DataVerse repository and the Federated Research Data Repository (FRDR). DataVerse is an open source data repository where Dalhousie researchers can deposit and share data. The DataVerse will hold data files, documentation, and descriptive metadata. FRDR is a part of that platform to support the preservation and sharing of Canadian research data.

By entering data into the Dalhousie DataVerse repository and FRDR, the research can be shared with a wide audience of interested parties and researchers. The team also recommends publishing the data, analysis, and findings to provide the widest audience.

Preservation-friendly File Formats
Ensure that preservation-friendly formats are used in long-term storage (.txt and .csv). To guarantee that information is anonymized, remove any identifying information from interview data. This includes direct identifiers, such as names, employment numbers, or place of employment; and indirect identifiers, such as gender, salary or age.
Include supporting documentation: describe study, method, metadata and backup processes.

SHARING AND REUSE

Form of Data to be Shared
The Principal Investigator may decide what form of data to share, and to base the level of data sharing upon the mandates of the funding body, the CIHR. JDC recommends not sharing the raw data to preserve anonymity of participants; instead, share data with identifying features removed or coded.

End-user Licensing
For the data collected through textual analysis and interviews, the recommendation is to obtain the following licensing for the data that will be released. The suggested license will ensure the Principal Investigator's rights as the original creator and ensure the data is used in the way they intended.

License: Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0)

This license allows others to share and adapt the work, but appropriate credit must be given, and any changes must be indicated. The data may not be used commercially, and if anyone builds on the data, they must use the same license. The open datasets that are used could have varying licenses, which means that the end-user license will depend on the individual dataset. The original licensing of the open data sources must be identified and maintained.

RESPONSIBILITIES AND RESOURCES

Project Data Management Responsibility
Primarily, Professor Green will be responsible for carrying out the data management plan and will hold responsibility for any modifications to the plan. He also will manage the data after the conclusion of the research project. As the project grows, it may be beneficial to designate a graduate student to maintain the plan. The student would be fully trained to learn the backup methods, metadata, etc.

In the unlikely event of a change in Principal Investigator, designate a co-investigator in the department (a trusted colleague) to obtain access.

Data Management Plan Resources and Cost

With CIHR funding, the major costs are the Box storage -- the total for three users is $252/year, or $756/year for the Business package. Colectica software has a free version and offers $12/month plan or a one-time $49 plan, each with more options. The purchase of a new secure external hard drive will cost between $50 to $150, with the addition of encryption software to ensure security.

ETHICS AND LEGAL COMPLIANCE

Security of Sensitive Data

Once interviews are conducted and the applicable data is extracted into the recommended template, the researcher will have two options for the identifying data:

1. Destroy the MP3 files and any documentation that identifies the individuals beyond the participants' coded identifiers to ensure ongoing protection of the identity of the participants.
2. JDC strongly recommends contacting the Dalhousie University Research Ethics Board to determine the protocol for storing sensitive information. If Professor Green decides to keep the raw data or the interview recordings with identifiers included, JDC recommends the purchase of an additional encrypted external hard drive for its storage. As it contains sensitive raw data, the device should only be accessible to Professor Green and stored in a locked, secure area. Further, it would be recommended that additional safeguards such as password protection be used.

Secondary Uses of Sensitive Data

A consent form will provide participants with information about how data will be shared and note that their data will be anonymized.

Legal, Ethical, and Intellectual Property Issues

The primary researcher must adhere to and pay close attention to the licensing of the public domain data that is gathered, as it will vary according to source. This will help to ensure that data is being used, modified, and shared appropriately in accordance with the appropriate licensing.

Data Management Plan

Project Name: Professor Pinkerton

Plan created by JACK Data Consultants

Principal Investigator of the Research: Professor Pinkerton

DATA COLLECTION

The project entails collecting data from internal and external sources, and then inputting the data into Excel. The types and formats of data that will be collected will vary according to the source. The types of data that will be created are numeric and text and should be saved in .xml or .csv format, which are open file formats that are widely used in the research community and will enable efficient data sharing and long-term access.

File Naming Conventions

JACK Data Consultants (JDC) recommends that the project be structured by distinguishing files based upon their origin (external versus internal), or by subject (student data, job description, etc.). Before proceeding with management and storage plans, it will be necessary to assess the structure of the files and to discuss with Neil Gaiman the current file structure and naming convention being utilized. For optimal navigability, an agreed upon convention should be established and utilized by both parties.

Files should be named using a standardized format. Records will be renamed so that the subject is clear, allowing for ease of retrieval. When adding dates use the standard YYYYMMDD format. If there is more than one copy of the same file or an updated dataset exists, undertake a careful assessment of whether it is necessary to keep the files. JDC recommends discarding any redundant or duplicate files to reduce the unnecessary use of hard drive space. If it remains necessary for the Primary Investigator to maintain multiple versions of the file, then clear indication of version number in a standardized format should be included in the file name (e.g., V01, V02) to eliminate confusion.

Assessing content, sorting, and renaming will be a lengthy process. The Principal Investigator should reduce the number of files first before committing to the renaming and uploading process.

DOCUMENTATION AND METADATA

Documentation

Documentation for all datasets can be inserted into each Excel file; this will ensure consistency and completeness. Documentation ensures that all information about a dataset is captured and is directly connected to the dataset to which it pertains. This includes the creator, context of dataset, revision history, source of data, and licensing information.

Documentation for the external data (from other researchers, various governments, and private corporations) that is kept as local copies can have documentation added. This includes: the

source of retrieval, the original type and format, and the method of transforming it from one format to another. If known, elements about the variables used, details of the study, and creation of the data should also be included.

Metadata Standards
Use a metadata standard for all the data documentation to ensure that it is navigable. To provide consistent metadata to the Excel spreadsheets, use a tool like Colectica (https://www.colectica.com/software/colecticaforexcel/) which allows for metadata documentation input directly to Excel and imbeds the metadata into each file. It includes the metadata elements: Title, Creator, Description, Data type, Variable descriptions. The documentation is exportable to XML. There is a free version of Colectica, with the option for upgrades for more features.

At minimum, use the following metadata elements, whether the datasets are internally or externally sourced:

- Title
- Subject/description
- Author
- Source
- Category
- Comments

Additional metadata elements can be added as needed to make the data retrieval process as efficient as possible.

STORAGE AND BACKUP

Storage Requirements
Professor Pinkerton currently has approximately 60 GB of data and will continue to collect datasets at a high rate. A storage system that accommodates the 60 GB and can grow and provide storage for as long as needed will therefore be suitable for the project.

Data Storage and Backup Plans
Standard practice is to store data in three separate places in the event of natural disaster, file corruption, or human error. It is also important to store data in multiple formats. Therefore, it is recommended that the research data be stored in three locations to avoid the loss of important research.

1. **Data.world.** Data.world (http://data.world) is a collaborative cloud platform used for organizing and sharing data. Using Data.world would eliminate the need to store data on a laptop and would enable the creation of datasets in the program itself instead of storing data in Excel files, should this appeal to Professor Pinkerton. Data.world allows the user to search open datasets and projects, share privately, create visualizations, and to tag data with rich metadata. Data.world will make sharing datasets much simpler.

2. **Dalhousie OneDrive.** OneDrive, which the Principal Investigator has access to as a Dalhousie faculty member, would be a suitable second cloud location to store and backup the Excel files.
3. **External hard drive.** JDC recommends storing a third copy of the data on an external hard drive that is kept in a secure off-site location. Although it is good practice to perform backups after every change made to data, JDC recommends creating a backup schedule and ensuring that files are backed up weekly, biweekly, or monthly, depending upon frequency of data collection or change.

Collaboration and Access

Data.world enables secure data sharing and collaboration, which will eliminate the need to email spreadsheets back and forth with colleagues. Data.world also allows members to invite others to view or contribute to their data and manage access controls by making data public or private. Professor Pinkerton currently wants colleagues to obtain permission to view datasets but not be able to edit them. Access controls may be changed in the future to allow colleagues to contribute.

PRESERVATION

Long-term Preservation and Access

With the recommended subscription to Data.world, collected research and datasets will be protected and maintained through the software's built-in monitoring practices. By utilizing Amazon web services and Google Cloud storage to house data, multiple safeguards and encryption procedures protect customer data on Data.world. Additionally, a variety of granular access controls allow the primary data contributor to determine who may view which data. JDC recommends granting the postdoctoral fellow full access, while allowing varying degrees of access to others who wish to view the acquired datasets.

Any research that is designated by the primary data contributor as open will be available online through Data.world for viewing, collaboration, and sharing. Data.world adheres to the findable, accessible, interoperable and reusable principles of research data, but offers the subscriber secure sharing features. The primary subscriber is in control and may grant access as deemed appropriate.

Preservation-friendly File Formats

Use a standardized file naming system to facilitate easier navigation of files and maintain use of .csv and .xml file formats. This will facilitate the easiest retrieval of data across all recommended platforms (Data.world, OneDrive, and the external hard drive). Further, Data.world offers a document keyword or tagging system to create more robust metadata descriptions and allows for fellow researchers to find data by subject or topic.

As 95% of Professor Pinkerton's data is open source, issues of anonymity are not a concern. However, if the student files or entry level job descriptions are kept as datasets, then a coding system should be employed to protect any names or designators of identity within the research

data. Once coding occurs, JDC recommends two options: Either to destroy the original raw data which contains any identifiers, or to contact the Dalhousie University Research Ethics Board to determine the protocol for storing sensitive information. If the Principal Investigator keeps the raw data or the interview recordings with identifiers included, JDC recommends the purchase of an additional encrypted external hard drive for its storage.

SHARING AND REUSE

Form of Data to be Shared

The recommendation is to share the raw data that is open data through Data.world. The Principal Investigator has full control over shared data.

End-user Licensing

The suggested license will ensure the Principal Investigator's rights as the original creator and ensure the data is used in the way they intended. JDC recommends using the *Attribution-NonCommercial-ShareAlike 4.0 International* license, which allows others to share and adapt the data, provided they give appropriate credit and indicate any changes. Additionally, this license prevents the data from being used commercially and if anyone builds on the data they must use the same license.

The open datasets that are used could have varying licenses, which means that the end-user license will depend on the individual dataset. The original licensing of the open data sources must be identified and maintained.

Publication of Data

JDC recommends using Data.world to share the datasets collected.

RESPONSIBILITIES AND RESOURCES

Project Data Management Responsibility

The client, Professor Pinkerton, and the postdoctoral fellow, Neil Gaiman, will be responsible for managing the data and following the structure of the data management plan outlined by JDC. We recommend that Gaiman's initial task involve renaming and uploading files to the Data.world repository, and that this online software remain the primary cloud storage system as data collection proceeds. Gaiman should also ensure that each dataset has proper metadata added. Data.world has step-by-step tutorials to assist with data uploading, and a user-friendly interface. JDC will guide the client through the initial phases of the data uploading process to ensure ease of software use and that best practices are established.

To ensure continued access to valuable research files and datasets, the Principal Investigator should allow the postdoctoral fellow or another trusted colleague to have full access to the Data.world online repository as well as the secured external hard drive. This will provide an additional safeguard and ensure that the data remains available to fellow researchers in the unlikely event of a change in project personnel.

Data Management Plan Resources and Cost
The Data.world package JDC recommends will cost $50 USD/month. This would be $600 USD annually, which would be between $750-800 CDN per year at current (April 2018) conversion rates. This plan allows up to five members, unlimited projects and datasets, and unlimited integrations. Colectica software has a free version and offers a $12/month plan or a one-time $49 plan, each with more options available than offered by the free version. An external hard drive would range between $25 and $100 CDN.

ETHICS AND LEGAL COMPLIANCE

Security of Sensitive Data
Should the Principal Investigator decide to retain student files or entry level job descriptions as datasets, then a coding system should be employed to protect any names or designators of identity within the research data. Once coding occurs, JDC recommends destroying the original raw data which contains any identifiers or securing the sensitive data onto an encrypted hard drive which will be kept in a secure off-site location.

Once the student and job description data is coded and all identifiers have been removed, there should be no concern with sharing the datasets via the online Data.world repository for secondary uses, should the Principal Investigator wish to share this information.

Legal, Ethical, and Intellectual Property Issues
JDC recommends undertaking stringent citation practices to ensure that the intellectual property of the creators of the data is respected, and to protect those who seek these datasets from mistakenly attributing the wrong person to the work. Following JDC's recommended use of a template to create metadata standards will protect the Principal Investigator and ensure that consistent records are kept regarding the source of each dataset.

Data Management Plan

Project Name: Professor Chartreuse

Plan created by JACK Data Consultants

Principal Investigator of the Research: Professor Chartreuse

DATA COLLECTION

The project will entail collecting data mainly from public domain resources and storing it in Excel files. The type of data that will be collected and created will be text and numeric. Data being stored in excel files should be saved in .xml or .csv format, and textual data should be saved in .txt format. These formats will allow for the data to be easily shared, accessed, and re-used by the researcher and anyone else who is provided access to the data.

File Organization, Naming, and Structure

JACK Data Consultants (JDC) have created an Entity Relationship Diagram (ER Diagram) that provides a visualization of how the data should be organized. This diagram is located at the end of this document.

Files should be named according to a standard naming system to allow for enhanced document searching and retrieval. By utilizing the design suggested in the ER diagram, issues around which type of file contains which type of data should be avoided. This design will allow for searching by keyword, date published, date retrieved, author, format, etc. This system will elevate the organizational structure of the data and allow for easier retrieval of the desired files.

DOCUMENTATION AND METADATA

Documentation for all Excel datasets should be inserted directly into each Excel file, which will ensure that all information about a dataset is captured and is directly connected to the dataset to which it pertains. JDC recommends using the following metadata standard to ensure metadata consistency among all Excel files.

Metadata Standards

A metadata standard for all of the data documentation aids in navigation and consistency. This will ensure that any data collected by others will follow the same labelling and metadata practices and allow for easy organization of the valuable information being collected and shared.

To provide consistent metadata to the Excel spreadsheets, use a tool like Colectica (https://www.colectica.com/software/colecticaforexcel/) which allows for metadata documentation input directly in Excel and imbeds the metadata into each file. It includes the metadata elements: Title, Creator, Description, Data type, Variable descriptions. The documentation is exportable to XML. There is a free version of Colectica, with the option for upgrades for more features.

Colectica is exportable to the DDI metadata standard for the social science research being conducted by the Principal Investigator. (http://www.ddialliance.org/training/why-use-ddi) This standard is machine-readable and interoperable, encourages comprehensive data description, and enables the reuse of metadata.

STORAGE AND BACKUP

Anticipated Storage Requirements

Professor Chartreuse currently has 20 GB of data that she has collected for four years. If data collection continues at this same rate, the total amount of data could increase to 70 GB within the next 10 years. Additionally, given that Professor Chartreuse is looking to expand his personal research, a storage system with upwards of 70 GB would be most suitable. Professor Chartreuse will also be able to store his data for as long as he wishes.

Data Storage and Backup Plans

Standard data practice is to store data in three separate places in the event of natural disaster, file corruption, or in the case of human error. It is also important to store data in multiple formats. Therefore, it is recommended to store the research data in three locations to avoid the loss of important research.

1. **Secure Cloud Storage. Box** is a secure cloud storage service. The Personal Pro plan will allow an individual research to share their data with other researchers. The fee for use is relatively low ($14 per month) and offers 100 GB of storage, which accommodates the needs for this project. Box also includes backup and version control features that will ensure that the data stored there is consistent and protected. Box provides *Box Zones*, so the client can preserve data inside Canada, preserving Canadian privacy requirements.

2. **Network Drive.** Another copy of the data will be kept on the Dalhousie Network Drive that Professor Chartreuse has access to as a member of the Dalhousie faculty.

3. **External hard drive.** JDC recommends storing a third copy of the data on an external hard drive that is kept in a secure off-site location.

Although it is good practice to perform backups after every change made to the data, JDC recommends creating a backup schedule. For instance, performing a backup operation each Friday prior to leaving the office will take very little time but will greatly reduce the likelihood that data will be lost. The Principal Investigator could determine the regular backup schedule based on the frequency of changes to the data (i.e., weekly, biweekly, or monthly). Backups should also be conducted after any substantial changes to the data.

Data Access and Modification

The access and sharing controls in Box will enable collaborators to access data (with permission from the Principal Investigator) without being able to modify it. The Principal Investigator will be able to share files of any size through individual links, or invite collaborators to view, access and search his entire collection of data.

PRESERVATION

Long-term Preservation and Access

JDC recommends a Canadian research data repository such as the Federated Research Data Repository (FRDR) (https://www.frdr.ca/docs/en/about/), which is currently in its BETA version. FRDR is a potential repository solution for researchers who do not currently have a practice in place. Researchers can upload files to FRDR while maintaining their metadata system, which eliminates having to re-enter metadata into the web form. FRDR also allows users to contribute up to 300GB of data, which should be enough to accommodate present data needs and accommodate any records and data acquired in the future. To ensure none of the important research data is lost, FRDR will also stored data in the long term with Archivematica (version 1.6.1), a stable and reliable web archival software.

Ensure that all files are in preservation-friendly formats for long-term storage (.xml, .csv, .txt).

SHARING AND REUSE

Processed and analyzed data will be shared in Excel spreadsheets. Documentation will be shared in its final form.

End-user License

Since data collection will be conducted using primarily public domain resources, the type of end-user license included will vary based on the specific data. It is recommended that the licensing of the sources from which data is collected be identified and maintained. Additionally, if it is not already specified within a license, it is recommended that you make the data available for non-commercial use only (e.g., a *non-commercial* condition can be added to Creative Commons licenses).

Publication of Data

Distribution on FRDR will help to ensure that the research community is aware of the data. The research may also be shared to a personal or institutional project website.

RESPONSIBILITIES AND RESOURCES

Data Management Plan Responsibility

The Principal Investigator will be responsible for managing the project's data during and after the project.

JDC recommends that the Principal Investigator ask a trusted colleague at the same institution to be co-investigator of the personal research data.

Data Management Plan Resources and Cost

The cost of the Box individual subscription is $168 for one year based on the monthly fee of $14. Colectica software has a free version and offers $12/month plan or a one-time $49 plan, each with more options available than offered by the free version. The purchase of a new

secure external hard drive will cost between $50 to $150, with the addition of encryption software to ensure security.

ETHICS AND LEGAL COMPLIANCE

Security of Sensitive Data

Since the data being collected is being retrieved from public domain resources, it is unlikely that it will include sensitive or personal information.
As mentioned previously, the suggested Box storage option includes settings that would allow for data to be accessible only to parties specified by the Principal Investigator should any sensitive data be included in the project in the future.

Legal, Ethical, and Intellectual Property Issues

JDC recommends adhering to and paying close attention to the licensing of the public domain data that is gathered, as it will vary according to source. This will help to ensure that the public data is being used and shared appropriately during and after the project. JDC has built into the data documentation plan a field for the specific licensing of the data that is acquired, which will make this information easily accessible and ensure that the Principal Investigator is protecting the rights of the original creators of the data. Additionally, the ER diagram for the research will be structured in such a way that will enable searching for records by license type. For example, all public domain records will be searchable by their public domain license.

Professor Chartreuse - Proposed Research Data Entity Relationship Diagram