DMP TITLE: PROF PINKERTON RESEARCH DATA MANAGEMENT PLAN

Project Name: Case 3 Prof Pinkerton
Principal Investigator: Prof Pinkerton
Institution: Dalhousie University

PART I: DATA COLLECTION

What types of data will you collect, create, link to, acquire and/or record?
The data will be collected is spreadsheet data. The spreadsheets include textual 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?
Now the data is in Excel format. We will take the text in Excel and pass it into TXT formant. The remaining data will be collected in CSV format, the recommended format for spreadsheets according to UBC Library.

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?
A convention will be used for name and version control.
The guidelines are as follows:
File names include last modified date and date is denoted in YYYYMMDD format.
File names include a short identifier or a summary of the content.
Use delimiter when necessary
Keep track of file version sequentially
Use simple folder hierarchy
The conventions will be documented for all team members.

PART II: DOCUMENTATION AND METADATA

What documentation will be needed for the data to be read and interpreted correctly in the future?
In order to make data to be read and interpreted correctly, a documentation will be prepared. The documentation will include:
Title
Principal Investigator (Creator in Dublin Core)
Project Description (Description in Dublin Core)
Each researcher's workload
Research methodology
Context of data collection
Methods of collecting data
Workflow of collecting data
Data source (Source in Dublin Core)
Data quality
Analysis method
Related publication
Related websites
Variable definitions
Format and file type of the data
Files structure
Permission on access

How will you make sure that documentation is created or captured consistently throughout your project?
We will determine individual role and workflow before data collection. We will consult research team on a regular basis. And we will keep progress report includes the last modified time and author.

If you are using a metadata standard and/or tools to document and describe your data, please list here.
Dublin Core metadata standard will be used when documenting the data.

PART III: 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?
It is certain that the data will be used in next multiple years, so the data will be stored for ten years. The average file size is 3.5MB and the total size of the data is estimated to be around 60GB. Considering the collection is rapidly growing, the anticipated storage is 200GB.
The way the data will be stored is the following:
Amazon Web Service: 200GB
Dalhousie University Dataverse: 200GB
USB Drive: 200GB
Laptop Hard Drive: 200GB

How and where will your data be stored and backed up during your research project?
To reduce the risk of losing data, the data will be stored both physically and electronically. Also, data will be deposited both online and offline considering the data includes open and non-open data. Depositing data both online and offline will help control access to the data. So, the data will be deposited in Dalhousie University Dataverse, Amazon Web Service, USB Drive, and Laptop Hard Drive. The USB Drive will be used as nonnetwork physical storage. The data will be stored daily to multiple locations during collection phase. More specially, the data will be stored daily on the laptop hard drive. At the same, three copies of data will be back up on a USB drive, Dalhousie University Dataverse, and Amazon Web Service Cloud separately. The project is long lasting. Every time Pinkerton uses the data to generate results, the results will be stored on the laptop hard drive.

How will the research team and other collaborators access, modify, and contribute data throughout the project?
The research team will access the data through Dalhousie University Dataverse and Amazon Web Service. They can download data form Dataverse and Amazon Web Service. Prof Pinkerton will assign roles to the research team members and collaborators in Dataverse, permitting various levels of access to the datasets. Prof Pinkerton will also add collaborators to Amazon Web Service. In this way, all the members can access, modify or contribute data. And it makes easier for Prof Pinkerton to keep track of changes. If any data is changed, the new version of data will be backed up.

PART IV: PRESERVATION

Where will you deposit your data for long-term preservation and access at the end of your research project?
At the end of the research project, the research data will be preserved for a long term. It is certain that Professor Pinkerton will reuse the data and research community will ask request to access the data. Considering the foreseeable use of the data and a large number of data requests, the data will be deposited in the place that is easy to access. The data will be deposited in Dalhousie Univeristy Dataverse and Amazon Web Service.

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.

To prepare data for preservation, we will do data cleaning including removing any identifiable information and handling missing data. We will also check errors to make sure the data will be preserved error-free. After data cleaning, the data will be converted into a preservation-friendly format. Excel files will be converted into CSV format for preservation and textual data will be converted into txt format. We will document any change and data lost during the format conversion. Metadata will be placed alongside the data to make data discoverable and reusable. The original research proposal and reports will be also deposited. References and local copies of reference sources will be deposited because it is uncertain whether external sources will provide data in the future. After those steps, the data will be ready for preservation.

PART V: SHARING AND REUSE

What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).
Raw data can be used by the research community as data sources of researchers. Processed data is the data ready for analysis, so sharing processed data will help other researchers effectively. Therefore, raw data and processed data will be shared with other researchers and the public. The data is categorized into open data and non-open data. No matter raw data and process data, only open data will be shared with the public and the sharing of non-open data is decided by Prof Pinkerton. In other words, the data will be sharing is raw data and processed data in open data.

Have you considered what type of end-user license to include with your data?
For open data, the license Creative Commons CC0 will be used in sharing of data. Professor is willing to share data and she is too busy with replying data requests. Therefore, there will be no restriction on access to data chosen to be shared. CC0 covers the copyright of shared data. By using CC0 Prof Pinkerton will dedicate the data to the public domain. CC0 will not restrict who can use the data. Prof Pinkerton will not claim copyright in the data and others can freely reuse the data.
In terms of non-open data, Prof Pinkerton is the intellectual property right owner of the data and bespoke licenses will be used. The sharing of non-open data depends on the case and Prof Pinkerton holds a record of who can access to non-open data. So bespoke licenses are suitable. Prof Pinkerton is suggested to seek guidance from legal departments to create an End User License for non-open data and a Special License for Sensitive data.

What steps will be taken to help the research community know that your data exists?
The following strategies will help the research community to find data:
-data is available on Amazon Web Service. Researchers only need to sign up an AWS account and then they will have free access to the data.
-Dataverse will assign a DOI to datasets.
-data will be cited in publications. The data will be referenced by a data access statement that provides the URL and identifier. So, the research community can access the data by the URL and identifier.
-metadata deposited with the data will help researchers to discover the data.

PART VI: 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.
Prof Pinkerton is the principal investigator(PI) and her fellow researchers will be CO-PIs. The principal investigator will have the overall responsibility for the research data management. The roles of data management team are as follows:
Part 1 data collection (Julie Timm)
Collecting data
File format
Conventions
Part 2 data storage and backup (Roberto Abarca Mero)
Back up
Version control

Progress report
Part 3 data sharing (Kiranteja Kolli)
Data protection

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?
If the principal investigator, Prof Pinkerton, leaves the project, Prof Pinkerton should notify co-PIs and the research team one month earlier before she leaves. Prof Pinkerton should look for another researcher to take over the work. If one of the team members will leave the project, Prof Pinkerton needs to select a new member. Current progress of the project and uncompleted work will be documented during personnel change.

What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?
The data management plan requires Amazon Web Service cloud. Prof Pinkerton is suggested to use pay-as-you-go payment approach considering the growth of data collection. According to AWS pricing policy, the first 50TB monthly fee is $0.025 per GB. If when the size of data grows to more than 50TB, Prof Pinkerton can purchase next 450TB and the monthly fee is $0.024 per GB. Current data collection will cost $1.5. The anticipated storage is 200GB and the cost will be $5. Prof Pinkerton might assign a research team member or hire an IT staff to be responsible for data storage and sharing.

PART VII: 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?
The research data might include sensitive data like human-related data. Sensitive data will not be saved on a laptop that connects to the network. Data cleaning will be done before saving data in Amazon Web Service Cloud, removing sensitive data. Sensitive data will be only saved on the non-network USB drive. The USB drive that will not be shared casually. The access to sensitive data will be limited and only Professor Pinkerton can decide who to access the data. Before sharing sensitive data, consent from related parties is needed. Sensitive data will not be shared by emails or cloud sharing services.

If applicable, what strategies will you undertake to address secondary uses of sensitive data?
A consent statement will be prepared. The use of sensitive data will be clarified in the statement. Other researchers will not be allowed to use sensitive data for commercial use. Other researchers will not be allowed to use the data to identify, contact or locate a single person.

How will you manage legal, ethical, and intellectual property issues?
CC0 is used to include open data. Using CC0 means surrendering the copyright of the data. There might be little legal or intellectual property issues around open data. The intellectual property right owner of nonopen data is Prof Pinkerton. Prof Pinkerton will use End User licenses to include her data. Researchers can access to nonopen data only after obtaining consent from Prof Pinkerton. A new type of licensing will be used for research results.

Prof Pinkerton has a legal duty of confidentiality if the data includes any confidential data. For confidential and personal data, the consent from related people to share such data is required. The research team will consult legal departments early in the research. The research team will pay attention to access control and data protection.