# My plan (Portage Template)

## DMP title: Pinkerton

### Admin Details

**Project Name:** Prof Pinkerton Research Data Management Plan

**Institution:** Dalhousie University

### 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 spreadsheet includes some text.

#### 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.

### Documentation and Metadata

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

the documentation will include:

the context of data collection

researcher

each researcher's workload

research methodology

analysis method

methods of collecting data

workflow of collecting data

data source

related publication

related websites

variable definitions

format and file type of the data

data quality

files structure

permission

#### 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.

### 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 median size is about 1000 rows of data and median size spreadsheet might be around 1MB. The total size of the data is estimated to exceed 20GB. Considering the collection is rapidly growing, the anticipated storage is 200GB.

The way the data will be stored will be the following:

Amazon Web Service: 200GB

USB Drive : 200GB

Laptop Hard Drive: 200GB

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

The data will be stored both physically and electronially. The data will be stored daily to multiple locations during collection phase. More specially, the data will be stored daily to the laptop hard drive. At the same, two copies of data will be backuped in a USB drive and Amazon Web Service Cloud seperately.

The project is long lasting. Every time Pinkerton uses the data to generate results, the results will be stored in 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 to the data through Amazon Web Service. Researchers might use email or sharing service like Dropbox to exchange data. They can also use Github to upload and download files. If any data changed during exchange, the new version of data will be backed up.

### 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 the large number of data requests, the data will be placed in place that is easy to access. Permission on access is not needed. The data will be deposited in Amazon Web Service. Researchers can have free access to datasets hosted by 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 preservation friendly format. Excel files will be converted into csv formant 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. Metadata includes keywords about data sets and the relationship between different data sets. 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.

### 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 research community as data sources of researches. 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.

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

The license Creative common’s (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. Using CC0 will put the data in the public domain. CC0 will not restrict on who can use the data. Prof Pinkerton will not claim copyright in the data and others can freely reuse the 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.

-data will be cited in publications. The data will be referenced by a data access statement that provide 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.

### 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 follow researchers will be CO-PIs. The principal investigator will have the overall responsbility for the research data management. The roles of data management team are as follows:

Part 1 data collection (Julie)

Collecting data

File format

Conventions

Part 2 data storage and backup (Roberto)

Back up

Version control

Progress report

Part 3 data sharing (Kiran)

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 team members will leave the project, Prof Pinkerton needs to select a new member. Current progress of the project and uncomplated 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.

### 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 in laptop that connects to network. Data cleaning will be done before saving data in Amazon Web Service Cloud, removing sensitive data before backup and preservation. Sensitive data will be only saved on a non-network physical storage, a hard drive. The hard drive that will not be shared casually, in other words, the access to the hard drive will be limited and only Professor Pinkerton can decide who to access data stored in hard drives. 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?

The data will be in public domain. Using CC0 means surrending the copyright of the data.