# My plan (Portage Template)

## DMP title

### Admin Details

**Project Name:** My plan- Case 3 (Portage Template)

**Institution:** Dalhousie University

### Data Collection

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

Excel spreadsheets are the only form of data concerning our purposes.

#### 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?

Although Professor Pinkerton prefers Excel she will also accept CSV, both of which, allow for data to be re-used, shared, and are useful for long-term access.

All formats of spreadsheets Professor Pinkerton prefer Excel but accepts CSV. All formats allow for data re-use, sharing and long-term access to the data.

#### 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?

Our team recommends using consistent logical structures for organizing data. Dates of spreadsheet are denoted YYYYMMDD (ex: 20180407). To help user searchability, our team recommends recording documents with a short unique identifier (e.g. the name of the project or its grant #) and using its file name to indicate its contents (e.g. Questionnaire or GrantProposal). However, projects usually involve multiple versions of projects over a long period of time. To maintain a high level of user searchability for the length of time a project takes, our team recommends employing a practical form of version control where document versions are tracked sequentially and uses \_ as a delimiter. Imagine Professor Pinkerton collecting job descriptions for entry-level positions in her field. To keep track of her files, she employs the following naming protocol - “file\_Jobdescription\_20180407\_v01”.

Another thing equally important to data organization is how you structure folders. It is vital for any organization to have a simple logical hierarchical design for storing folders and maintain its consistency by strictly enforcing its practices. Simple folder designs are easy and intuitive for users to navigate and have the additional bonus of taking less time to perform a system backup compared to needlessly complicated folder designs.

dates of all spreadsheets are better in YYYYMMDD format, such as 20180407. All documents use a short unique identifier (e.g. Project Name or Grant #) with a summary of content (e.g. Questionnaire or GrantProposal) as part of the file name, which can help people to search by file names. Moreover, Keep track of document versions either sequentially and Use \_ as delimiters can make files easier be searched. For example, Professor Pinkerton collected job descriptions for entry-level positions in her field. These documents can be named "file\_Job decription\_20180407\_v01". In addition, the designer should make folder hierarchies as simple as possible. Because the complex folder hierarchy is difficult to navigate, it provides more opportunities for commit errors. At the same time, the system backup may take longer.

### Documentation and Metadata

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

Our documentation encompasses Professor Pinkerton’s research, data-level descriptions, and any background information other researchers require to access the data. More specifically for Professor Pinkerton, our documentation consists of information about spreadsheets, descriptions of each spreadsheets, and instruction of how people can use those spreadsheets.

Potential other elements that should be recorded include: the research method used, data format and file type, the description of how the data was gathered and the methods of collection, and who performed the tasks in the project and with detailed notes for each task.

In general, good documentation includes information about research, data-level descriptions, and any other background information needed to make the data available to other researchers. In this case, there are three type of documentation should be included: information about spreadsheets, the descriptions for each spreadsheets, and instruction of how people can use those spreadsheets. Other elements may need to be recorded, including: the research method used, the format and file type of the data, the description of the data capture and collection methods, and who has performed the tasks in the project and performed detailed information for each task.

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

Over many years, Professor Pinkerton’s data has been collected from a high volume of highly diverse sources and these complex origins are reflected in the. To simplify this complicated content, Pinkerton only accepts Excel files mostly but occasionally accepts CSV.

According to the case data, Prof Pinkerton collects work-related data, such as her recorded student performance data for 12 years. At the same time, she also regularly saves interesting data sets for later use. She often downloads files of interest on the open data portal. At the same time, colleagues and friends also send her files. Since all the files will be Excel and CSV files, the format is determined. The contents of the documents need to be reviewed by the professor and later classified.

#### If you are using a metadata standard and/or tools to document and describe your data, please list here.

Our team recommends using Dublin Core Metadata as a metadata standard.

The metadata standard I will choose is Dublin Core Metadata Initiative.

### 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?

Carefully organized on Professor Pinkerton’s laptop are 17,384 spreadsheets with row counts ranging from 1 to 750,000. Our research team estimates that for effective storage…

Since Prof Pinkerton has a total of 17,384 forms. The smallest file has only one row of data, and the largest file exceeds 750,000 rows of data. The common file size is almost 1,000 rows of data. After calculation, the storage range can be roughly.........

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

A backup strategy is important for negating losing data due to human error, natural disasters, and otherwise unforeseen events. Best practices of data management indicates having at least three saved copies of data consisting of one primary file and two backups stored in two different forms of media with one kept offsite. Forms of data storage include using removable, fixed or networked optical or magnetic media. To differentiate forms of media, a third copy of data can be stored on formats such as DVD, USB drive, desktop hard drive, etc.

A reasonable backup strategy can reduce the risk of data loss due to human error, natural disasters, or other unfortunate events.  
In general, data management has at least three of your data, and the copy is stored on two different media. For example, keep a backup copy offsite  
Data can be stored using removable, fixed or networked optical or magnetic media. In this case, Prof Pinkerton kept a copy of the forms on his computer, and Neil Gaiman would create a cloud folder for re-storage. It is recommended to seek a third time for storage on another device, which can be a DVD, USB drive, desktop hard drive, etc.

#### How will the research team and other collaborators access, modify, and contribute data throughout the project?

A more secure strategy than relying on email to share files is to use a third-party commercial file sharing service such as Google Drive or Dropbox. Although there is no long term guarantee that these services will last, they can guarantee a degree of information security through limiting user rights to access, change, or contribute to data sets and can even control how many people can access the data.

In order to promote cooperation and ensure data security, necessary file sharing strategies need to be established. However, transferring data between locations or within a research team can be challenging for a data management infrastructure. In the case, Prof Pinkerton relies on e-mail to communicate with colleagues, which is not a powerful or secure solution. A more appropriate solution is to use third-party commercial file sharing services (such as Google Drive and Dropbox) to facilitate file exchange. However, they are not necessarily permanent or safe in the long run.For this issue, they can guarantee a certain degree of information security by limiting user rights or controlling the number of visitors.

### Preservation

#### Where will you deposit your data for long-term preservation and access at the end of your research project?

To create accessibility for Professor Pinkerton’s data, our team recommends dividing files into several categories according to Pinkerton’s data collection habits (e.g. job descriptions, student performance data, open data portal, etc). Each of these type of data are set to a file retention period table. so that when it expires it will perform any relevant processing methods required to permanently retain it.

For data retention issues, the earlier the designer is planning for data retention, the more beneficial it is for long-term data retention. For Prof Pinkerton's data collection habits, it can be divided into several categories according to the contents of the file, such as job descriptions, student performance data, open data portals and so on. Each different type of file needs to be set a file retention period table. When it expires, it needs to perform relevant processing methods, such as permanently retaining or destroying it. Because professors are fond of forms, many data will not be eventually destroyed. Therefore, all data can be managed based on the file retention period. However, after the documents expire, the professor can transfer all files to the cloud folder for permanent preservation.

#### 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.

Sometimes information is lost when converting a file to another format. To prevent this from happening, both a source file and the newly create file need to be recorded before any conversion occurs. This should fit well with Professor Pinkerton’s practices of using most Excel and CSV files in a carefully organized folder system. New files must ensure that all data is anonymized, error-free, and records the data into the recommended format to minimize data loss.

Since the information may lose information when the file is converted from one format to another, the file format of the record should be determined at the time of recording. In this case, the file format was determined for Excel and CSV. For a previously existing file, if the file needs to be converted, both the source file and the new file need to be recorded to reduce the possibility of information loss. For new files, you need to ensure that the data that you choose to keep or share is anonymous, error-free, and convert the data to a recommended format to minimize the risk of data loss.

### Sharing and Reuse

#### What data will you be sharing and in what form? (e.g. raw, processed, analyzed, final).

Raw, processed, analyzed, and final data are to be shared. Because Professor Pinkerton has 95% of her data coming from external sources, most of this data consists of final data. Therefore, most of Pinkerton’s data is ready to be shared. However, Pinkerton has a large collection of raw student performance data collected over the course of 12 years. This data must be anonymized to ensure privacy is maintained.

There are four types of statistical data: raw data, processed data, analytical data, and final data. According to the case, the professor should classify all the information she has. According to the article, 95% of the Prof Pinkerton forms have external sources, such as data from other researchers and the government. If these file information is final data, that is, data that has already been processed and can be safely saved. Then it can be set as shared data. If it is raw data, such as 12 years of student performance data, it needs to be protected for privacy considerations.

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

Our team designed this data management approach to be able to be shared as widely as possible therefore we will use public domain licenses. This license means most of the data can be used by anyone for any reason anywhere. Our team also wants to make it clear that choosing this license requires careful attention to the control and management of databases to prevent them from being harmed by unfair competition.

This cloud folder was created to use data as freely as possible and to devote work to the public domain. So as a knowledge sharing organization and open data sharing organization will use public domain licenses. In this way, most of the data in this database can be used by anyone with any tool at any destination. The use of public domain licenses requires attention to the control and management of databases to prevent them from being harmed by unfair competition.

### 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.

This database will be created by Neil Gaiman and will be managed by Pofessor Pinkerton. We trust that having over a decade of experience she has developed a methodology capable of maintaining such an impressive collection.

#### 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?

Before leaving Pinkerton’s projets, all parties responsible for data management activities will create a formula for anyone stepping into their position including: best data management practices, naming formulas, folder hierarchy strategies, and protocols for passing on information to people new to their data management positions.

Before the two responsible persons leave, they need to formulate the strategy for inheriting these data. This includes describing the process that the responsible person should follow when he leaves the project.

#### What resources will you require to implement your data management plan? What do you estimate the overall cost for data management to be?

In addition to previously covered financial costs, management costs include: technical costs of data management, training costs, file storage and backup, and contributions from non-project personnel, all of which, are broken into long-term and short-term costs.

The cost of this one data folder includes the data management costs incurred during the establishment of the project and the costs required for long-term support of the data after the project is completed. According to the length of data folder management, management costs are divided into long-term and short-term. Specific cost items include technical costs of data management, training costs, file storage and backup, and contributions from non-project personnel.