# My plan (Portage Template)

# Chief Writer: Li Liu b00639704 (Edited by: brian Jenkins)

## 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.

#### 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”.

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.

### 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.

.

#### 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 predetermined.

Since this cloud folder was created by the professor’s assistant, Neil Gaiman, it is important to incorporate his document management skills into the design of folder organization. In addition, Gaiman’s close proximity to Pinkerton’s research will allow for more intuitive folder organization other than a typical hierarchical design. For example, colleague emails could be arranged alphabetically by name and include a list of person’s frequently communicated with. 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.

### 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 chose to err on the side of caution and project for maximum data storage needs of 12.13TB of data. The breadth of Pinkerton’s data sets means that the length of storage time is indefinitely.

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

A backup strategy is important for negating data loss due to human error, natural disasters, and otherwise unforeseen events. Best practices of data management recommends 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. Different forms of media, such as CDs or DVDs, are convenient for daily input of data. However, it is more convenient to back up large amounts of data on a network hard disk. To prevent data loss, backing up all data after an update is recommended.

#### 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 schedule. Before a file classification plan is enacted, Gaiman should help classify all files entering into storage with retention indicators such as activation and deactivation times. This is important to prevent files unnecessarily occupying space since once a file has run out of use towards its project it can be transferred to the cloud database. This is also relevant for files that have been found to be in error (e.g copyright issues, or original files have been updated). By removing these unnecessary files the important one’s can be accessed more easily. Although not as secure a long-term strategy compared to localized options, cloud storage has convenient information sharing capabilities. However, all data should always be saved on multiple devices.

#### 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 the 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.

### 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 will not be shared and must be anonymized to ensure privacy is maintained.

#### 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.

### 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 totaling $17,176.08.