Group Project

**Professor Pinkerton**

Introduction and Recommendation

Professor Pinkerton has a very large, and ever growing, collection of data that is used by many people. Some of her files contain open data and she would like everyone to have access to them, while some contain sensitive information with access awarded on a case-by-case basis. This data management plan attempts to provide a way for Professor Pinkerton to easily organize her data sets so that they can be used in the future, either by herself or other individuals.

Given the vast number of spreadsheets that Professor Pinkerton currently has (and the amount that she is likely to collect in the future), it is our recommendation that a database management system is set up to organize this data. Contained within this DBMS will be the metadata about all of the different spreadsheets that she has collected. This DBMS will help Professor Pinkerton keep a record of all of the data sets that she has in an organized fashion, as well as help her locate them easily when needed. It will primarily operate as a database with information about all of the data sets that Professor Pinkerton collects. This DBMS should be available online so that anyone with data requests can easily find what data Professor Pinkerton has. The DBMS should also include links to where the open data sets can be downloaded so that they can easily be accessed.

We are recommending that all of the spreadsheets should be hosted on a drive, with a hard disk back up. The files will have to be separated in to two different categories (with the category of data being noted in the metadata, and therefore in the DBMS). The first category of spreadsheet is for open data, which anyone can have access to. The second category of data is for files that Professor Pinkerton would like to control who has access to.

The spreadsheets that are in the first category will be on a drive that is open and accessible to anyone, while the spreadsheets that are in the second category will be on a separate drive, and each of these files will be protected so that only people that Professor Pinkerton has given permission to can access them. People will have to request permission to access these spreadsheets. Both of these drives will be protected so that no one, other than Professor Pinkerton and potentially her postdoc students (at her discretion) can upload or modify files that are on the drive. This will prevent any changes from being made to the raw data sets that are a part of her collection

Necessary Data

There is no new data that Professor Pinkerton actually needs to collect as a part of this data management plan. Rather, as her collection of data increases and she gets more and more spreadsheets, these new sheets will be integrated into her collection based on the database management system that gets set up. The metadata will need to be entered into the DBMS as the spreadsheets are added to the collection, and will need to be uploaded to the appropriate drive location (depending on whether the data is open or not), and backed up on the local backup.

Professor Pinkerton does need to locate and organize the metadata about the different data sets. This is all information that she currently has, it will just need to be compiled into the DBMS.

Licensing Information

The licensing information for each of the spreadsheets needs to be contained in the metadata, and therefore noted in the DBMS. Each spreadsheet will have its own licensing requirements, and will therefore need to be tagged as either copyright, copyleft, or creative commons (BY, NC, ND or SA). Whichever licensing is on the data sets when Professor Pinkerton obtains them, need to be included in the DBMS.

Necessary Facilities/Equipment and Costs for Data Storage

It is our recommendation that all of the data be stored in two locations. The first location is on the university cloud that is used at Dalhousie University. The second location is to have local backups of the data on the hardware where the files are currently stored. As a result, no new facilities or equipment are required to organize and maintain the data

Because of the vast amount of data that Professor Pinkerton currently has, and the rate at which she is continuing to compile data, more storage may be necessary in the future. This would require both an expansion of the cloud that Dalhousie uses, as well as more hard drive space to store the data locally. Because Dalhousie University already has both an online storage drive, and hardware for backing up this information, no additional costs will be incurred to store this data.

We also recommend that the DBMS is created using a program that makes use of a search query language (SQL). One program that she should consider using is SQLite, which can be downloaded for free, and thus again resulting in no additional costs. This DBMS should be backed up both on the cloud and locally, similar to the other data sets.

Ownership and Access to Data

The DBMS would be owned by Professor Pinkerton, however should be available to anyone that is looking for data sets. Professor Pinkerton can provide a link to the DBMS to anyone looking for data so they can locate what they are looking for, without having to request the data sets from her by email (except for the protected sheets).

As mentioned before, the spreadsheets that Professor Pinkerton has will need to be separated into two different categories. The different categories will have different access and ownership characteristics. For data that is considered to be open, the spreadsheets should be made publicly available on the Dalhousie University online drive. This will reduce the amount of requests that Professor Pinkerton will receive to send out her data sets as people can download these data sets themselves.

For the second category of data sets that are not considered to be open, the access will be issued on a case-by-case basis. These will be stored on a location of the drive that is not shared. For someone to access one of the data sets, they will need to seek permission from Professor Pinkerton, and she will send them the link to access the specific data. This will also allow her to turn off link sharing of any of the files when she no longer wishes to share the file with others.

Post-Project Data Value

By setting up this DBMS, significant value is obviously created. Once the data sets are hosted online and interested parties have access to the DBMS, the data sets will be much easier to find for interested parties. Additionally, it will be easy to add new metadata for the DBMS as new data sets are collected. It will also reduce the number of data requests that Professor Pinkerton will receive, as interested parties will be able to download the data that they need (as long as it is tagged as open data). This DBMS also helps Professor Pinkerton comply with the FAIR data principles, adding further value. The DBMS will help with all aspects of the FAIR principles (findable, accessible, interoperable and reusable).

Metadata Requirements

The inclusion of accurate and thorough metadata is critical for maintaining the DBMS in an organized fashion. The metadata that needs to be included about each of the spreadsheets should be:

* Title
* Subject
* Size
* Original Author
* Licensing Information
* Whether the file is Open or has restricted access
* Who has access to the file
  + For files tagged as open, the default should be that anyone has access to the file
* Date
* Size

This data should be entered into the recommended SQLite program for each new data set that is collected. To metadata should be accurate and complete for each set to ensure that a correct record of all of the data sets that Professor Pinkerton owns are logged and can be accessed by others with ease.

Overall, by creating a DBMS, Professor Pinkerton can ensure that her data is organized and easy to use for both herself, and others in the future. By storing the data both on the Dalhousie University drive, and having local backups, Professor Pinkerton can be confident that she will have this data for the foreseeable future.