DATA MANAGEMENT PLAN - MANAGEMENT INFORMATION SYSTEMS AND OTHER RESEARCH TOPICS

Consultants: F.A.M.E DATA MANAGEMENT – for data that are **F**indable, **A**ccessible, **M**emorable, & **E**nduring. Team members: Jenny Farbstein, Lisa Nardecchia, Victoria Roberts, Carlye Stein

Principle Investigator: Professor Pinkerton

Funder(s): Not disclosed

Plan Details:
The following data management plan has been developed for Professor Pinkerton who is interested in open and non-open data relating to Management Information Systems and other research topics.

Elvira Mitraka
20180410

DATA COLLECTION

A variety of data are collected in different formats which are converted to Excel. Ninety-five percent (95%) of the data are from external sources, including open data portals, open government datasets and content contributed by other researchers. The content relates primarily to Information Managment Systems although there is some data Professor Pinkerton has generated from her own research that has also been converted to Excel. The average file size is 3.5MB, and total size of data collected to date is approximately 60GB.

Excel is not an open format, and is prone to conversion errors that are difficult to detect in larger spreadsheets, along with other issues that jeopordize data integrity during analysis. It is therefore suggested that all existing Excel documents be batch converted to comma-separated-values (.csv) files and that the use of Excel's default file extension be discontinued.

The .csv format is an appropriate selection as it is an open and compact format that will help to maintain future interoperability and longevity of the data. Other advantages of the format are that it is more robust for the purposes of analysis and reproducibility, and easily yields itself to version control. As the long-term goal is to use data to answer research questions, publish papers or to produce visualizations, converting to .csv is aligned with this goal and will help to assure both the accuracy of the final output as well as the long-term preservation thereof.

Data are downloaded from various sources. Other researchers send files directly which they think will be of interest. For added quality assurance, using an open source tool such as DataCleaner is being suggested. It is a data profiling engine for discovering and analyzing the quality of data. It will explose patterns as well as missing values and character sets along with other noteworthy characteristics in the datasets.

Data that are not complete, or that can neither be made complete nor cleaned, should be discarded.

DOCUMENTATION AND METADATA

Since the recommendation is that files are batch converted to .csv format, we suggest using CSV Engine (http://data.wu.ac.at/csvengine) to assist with the creation of metadata and documentation. This resource provides tools for cleaning, profiling and metadata generation according to web metadata specifications. We also recommend establishing an ‘about’ or ‘readme’ document describing the content and organizational structure of the files being uploaded and structured in the internal Cloud. This will ensure the research team can understand, identify, and locate files quickly and easily.

ETHICS AND LEGAL COMPLIANCE

To ensure ethical compliance for any future research, it is recommended that the guidelines established by the institutional Research Ethics Board be followed.

Copyright and Intellectual Property Rights (IP/IPR)

As there is currently no data ownership strategy in place, we recommend that the following files be copyrighted: the quantitative literature review and the collection of job descriptions and student performance information. In order to do this, the nature of the data and time that the files were created need to be documented and retained as metadata associated with each file. This provides ownership of the material and the proper license can be selected if and when the data are anonymized (where applicable) and if it is to be made accessible to others.

The licensing type for existing data being retained and shared needs to accompany the data and become part of its documentation. Open source formats and repositories should be recommended to colleagues who seek to share their datasets with the Professor Pinkerton. It is also helpful to create a publically accessible document outlining these recommendations. A follow-up process should be established to indicate that new data has been contributed. Once the location of the new data is documented, it is to be retrieved for cleaning and assesment, with only necessary content kept and attributed to the contributor if required.

STORAGE AND BACKUP

We recommend that the researcher utilize available institutional options, specifically OneDrive (dal.ca/dept/its/o365/services/onedrive.html) for storage and back up (done automatically). A Dalhousie University Dataverse account (dal.ca.libguides.com/rdm/daldataverse) should be acquired and used for storage and safekeeping, and the current practice of using a personal laptop discontinued as soon as an alternative is in place.

OneDrive should be used as the Professor's personal, private and primary storage area. Laptop data should be uploaded to OneDrive and purged from the source device. We suggest that only the Professor Pinkerton and their postdoctoral student Neil Gaiman gain access to OneDrive. Permissions can be granted at a file or folder level so the researcher can monitor how files have been accessed.

As the student performance data is highly sensitive, it is appropriate to have tighter controls in place for its storage and access. We suggest that a virtual machine is created with access limited to only the data owner. Student records will be stored in this location with weekly incremental backups in place and a monthly backup to OneDrive conducted for redundancy. If Neil Gaiman must have access, it must be to an anonymized version with only Professor Pinkerton having access to the key.

OneDrive has sufficient storage for the researchers' needs and offers reliable online access from any device. OneDrive offers accessible, secure and free storage and backup, making it a cost effective and easy-to-use tool to implement as part of this data management plan. Once OneDrive is set up and synced, we recommend that all laptop data be moved to OneDrive and removed from the source device once OneDrive content is confirmed.

Dataverse is a repository hosted on the university's own secure servers that allows users to search, download, and post data (and associated documentation and metadata) and is set in association with the university's Research Data Management Team (RDMT) who are available to assist throughhout the lifetime of the account. We recommend that administrator status be requested from the RDMT so that Professor Pinkerton has full control over account and access settings. This will allow them to assign the appropriate membership and permissions regarding what level of activity (read, write, modify, execute, delete) is applicable to persons with access.

Another advantage of Dataverse is that it offers version control; if editing privileges are granted it allows you to monitor what modifications are made and by whom. As Dataverse only supports certain formats including .csv, it is especially important that file conversion takes place as previously discussed. Dataverses enables the creation of sub-dataverses, which is useful when organizing the size and variety of data that will continue to be acquired.

Access and Security

Security concerns are most likley to apply to the collection of historical student performance data. For this reason, access to the folders that contain this data should be given only to the owner. Datasets hosted in Dataverse can have relatively lenient access controls applied to them. However, since preservation and integrity of the data is a key concern, modify and delete access will be limited to Professor Pinkerton, with additional permissions decided on a case by case basis.

SELECTION AND PRESERVATION

Data that are already stored in reputable repositories will be purged, with additional interaction taking place only through the existing access points.

As many of the saved datasets were collected several years ago, it recommended to review the downloaded files against the source data to ensure that: the most up-to-date and relevant source data are identified, and appropriate source access points are documented for future retrieval.

Student data will be preserved in its entirety and should be backed up in full every term. Incremental backups should also be taken on the server to ensure the most complete backup is available. We would recommend leveraging self-paced learning resources on DataOne's data lifecycle. For more information please see https://www.dataone.org/data-life-cycle

Long-term preservation plan for the datasets

As there is already an established Dataverse repository, long-term preservation of the data will be handled through Dataverse for the duration of the Professor's affiliation with the institution. If a change in affiliation is to occur, the datasets should either be transfered to the new institution's Dataverse instance or sent to the server currently used only for student performance data.

DATA SHARING

Both data being shared with the researcher, and data that is being sent externally at the request of interested parties will be shared using FileExchange. Other sharing and collaboration will be built into the permissions of the internal Cloud environment being built out.

RESPONSIBILITIES AND RESOURCES

Responsibility for data management

We recommend data management responsibilities be transferred to Neil Gaiman. Best practice overview and initial assistance will be provided by F.A.M.E Data Management and transitioned after the end of their responsibilities.

Resources required to deliver the data management plan

Information Architecture Resources
OneDrive
Dataverse
CSV Engine
Data Life Cycle from DataOne
Virtual Machine

Human Resources
Postdoctoral student(s)