INFO 6540

Elvira Mitraka

Samantha Adema, Brian Jenkins, Li Liu, JP Pineo

Group Assignment – Data Management Consultation Report

April 10th, 2018

Chief Writer: JP Pineo - B00522848 (Edited by Brian Jenkins - B00782701)

Group Assignment - Case 1

April 10th, 2018

Case 1: Professor Periwinkle

MY PLAN (PORTAGE TEMPLATE)

DATA COLLECTION

WHAT TYPES OF DATA WILL YOU COLLECT, CREATE, LINK TO, ACQUIRE AND/OR RECORD?

Raw spatial data, images, and text will be collected from: remotely-operated marine vehicles (ROMV), tags that are surgically implanted in captured and released animals, static sensor buoys that measure ocean conditions, and communication lines that passively listen for signals from animal tags. At a later date, data from the public will be connected as well.

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?

Collected data will be converted to NetCDF formatted files.

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?

For version control, our team suggests using GitKraken as it is considered especially user friendly. GitKraken's user friendly qualities extends to organization efforts as its design helps users understand where specific forms of data are being stored. The content files will be labeled according to what kind of content is being

used in them. For example, they will be labelled based upon specimen, location, date, and how and where it was collected from

(e.g. 20180410_seal_thunderbay_eartag)

DOCUMENTATION AND METADATA

WHAT DOCUMENTATION WILL BE NEEDED FOR THE DATA TO BE READ AND INTERPRETED CORRECTLY IN THE FUTURE?

We recommend using Dublin Core metadata, because it gives preliminary and necessary metadata to ensure interoperability while being easy to read. To ensure no important parts are overlooked, the data management team will need information about where and when the data was collected and the significance of each piece of data. This will help the data management team create effective and consistent metadata records in order to keep the most necessary information and increase ease of access.

HOW WILL YOU MAKE SURE THAT DOCUMENTATION IS CREATED OR CAPTURED CONSISTENTLY THROUGHOUT YOUR PROJECT?

After each important section of information is put into the metadata, our team will incorporate a secondary review process to ensure accuracy and consistency. For consistency, our team will create a standard method of document creation. The data managers will go through each documentation of data and create a new document that is needed for each one before creating an official document. Our team will use GitKraken to take advantage of its version control capabilities. Our team will upload each document as we work on them so that the changes from each version can be assessed for consistency with our created method for metadata.

IF YOU ARE USING A METADATA STANDARD AND/OR TOOLS TO DOCUMENT AND DESCRIBE YOUR DATA, PLEASE LIST HERE.

In part because it is machine readable, our team recommend using Dublin Core for a metadata standard.

Attached is a link to the different metadata elements that will be needed to create the metadata scheme for Dr. Periwinkle. http://dublincore.org/documents/dces/.

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?

We recommend moving Dr. Periwinkle's data from Dropbox to another cloud storage platform with greater storage capabilities. This data storage needs to be capable of expanding by 500MB a day and other cloud platforms are capable of this level of storage scalability. We recommend staying with a cloud storage method because Dr. Periwinkle's requires her data to be open to anyone who desires to use it. For all these reasons, we recommend Google Drive as an appropriate solution. Although it requires login credentials, which may hinder the open data goals of Dr. Periwinkle, our team recognizes that the added security of determining who has access to this data is a worthy pay-off.

In Google Drive there can be multiple files and folders shared within a master folder, this would make collaborating multiple people on the research team easier. It will also allow each researcher and any grad students to have their own folder

within the main research folder. Anyone can add their own data into a shareable file and have as many folders as needed for organizing different parts of their study.

This format is appropriate considering the high volume of student-collected data contributing to Dr. Periwinkle's collection. Because of the ever-changing nature of the relationship Dr. Periwinkle's students have to her research goals, putting data in a cloud will allow any data that is collected to remain in the cloud even after students potentially leave. In addition, Dr. Periwinkle and her staff need consistent access to this storage system regardless of location and employment.

HOW AND WHERE WILL YOUR DATA BE STORED AND BACKED UP DURING YOUR RESEARCH PROJECT?

Google Drive automatically backs up and saves all work. For additional back up, external hard drives will be used and updated once a week to include any new data and information. Each eternal hard drive will be 2TB and will be used to store all collected data and information as a failsafe in case the cloud storage experiences data loss.

HOW WILL THE RESEARCH TEAM AND OTHER COLLABORATORS ACCESS, MODIFY, AND CONTRIBUTE DATA THROUGHOUT THE PROJECT?

Collaborators can add or change aspects of their contributions in Google Drive. Google Drive automatically records who performed changes, what they changed and when, and makes this information available to everyone using it.

WHERE WILL YOU DEPOSIT YOUR DATA FOR LONG-TERM PRESERVATION AND ACCESS AT THE END OF YOUR RESEARCH PROJECT?

At the end of Dr. Periwinkle's study, all data will be given to the Department of Fisheries and Oceans Canada (DFO) where it will be kept for long term storage. DFO will receive all raw and analyzed data. DFO will grant access to whoever needs this data through a simple request process.

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.

For preservation purposes, all data will be converted into a text document, in case difficulty arises from technology not being able to read NetCDF format. These changes will be noted, and if there is any data lost in the conversion it will be noted, and the data that is missing will be noted as to what was lost. These changes will be noted because all files will be kept in NetCDF format in addition to a word document. By storing the data in two formats, research members can compare the two files for discrepancies that indicate data loss.

SHARING AND REUSE

WHAT DATA WILL YOU BE SHARING AND IN WHAT FORM? (E.G. RAW, PROCESSED, ANALYZED, FINAL).

Shared data includes all data forms (raw, processed, analyzed, and final) collected from ROMVs, tags that are surgically implanted in captured and released animals, static sensor buoys that measure ocean conditions, and communication lines that passively listen for signals from said animal tags.

HAVE YOU CONSIDERED WHAT TYPE OF END-USER LICENSE TO INCLUDE WITH YOUR DATA?

Our team recommends a simple Open Data Commons licenses, specifically Attribution License (OCD-By). This license allows all data collected into Dr. Periwinkle's database to be accessed by those who need it which is in line with her research goals.

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.

During the duration of the study Dr. Periwinkle's data will be stored with our research team. A small group of three people will be assigned to look after the data by ensuring it is kept up to date with the latest findings from Dr. Periwinkle's team and carefully organized and saved in appropriate files. Working closely with Dr. Periwinkle and her team will allow the data management team to understand the data being collected, and in turn, this will help organize it. Each of the three team members of data managers will have their own specific job on the team; one person will collect the data from Dr. Periwinkle and her team of researchers, one will organize it and upload it onto the new Google Drive files and onto the external hard drives, and the third person will ensure the consistency and completeness of each file. After the study is complete all data will be collected and delivered to DFO. The data will be accessible to the public and other marine wildlife research teams

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?

All data is added to Google Drive by Dr. Periwinkle's team of graduate students and our team of data managers will add this data to an external hard drive. Adding Periwinkle's data to an external hard drive is beneficial for when students inevitably leave Dr. Periwinkle's team, and decide they want to take their data with them, it can be retrieved and re-uploaded to Google Drive.

WHAT RESOURCES WILL YOU REQUIRE TO IMPLEMENT YOUR DATA MANAGEMENT PLAN? WHAT DO YOU ESTIMATE THE OVERALL COST FOR DATA MANAGEMENT TO BE?

Resources required include a group of 3 data managers, 4 2TB external hard drives, and a means of storing the hard drives offsite. An estimation of the costs for implementing this five-year data management plan is $907, 250.00.

ETHICS AND LEGAL COMPLIANCE

LEGAL, ETHICAL, AND INTELLECTUAL PROPERTY ISSUES?

Since Dr. Periwinkle has grad students collecting data for her, a contract should be created stating that the research and data they collect is the intellectual property of Dr. Periwinkle and they consent to allow others to use and store this collected data. This contract will ensure no legal discrepancies regarding intellectual property rights occur between Dr. Periwinkle and her student researchers.

References:

Question 6: 2018. Dublin Core Metadata Element Set, Version 1.1: Reference Description. Dublin Core Metadata Initiative. http://dublincore.org/documents/dces/

Question 17: 2018. Data Manager Salary (Canada). Pay Scale. https://www.payscale.com/research/CA/Job=Data_Manager/Salary

Chief Writer: Samatha Adema – B00736192 (Edited by Brian Jenkins – B00782701)

Group assignment - Case 2

April 10th, 2019

Case 2: Dr. Green Research Plan

MY PLAN (PORTAGE TEMPLATE)

ADMIN DETAILS

Project Name: Dr. Green Research Plan

Principal Investigator / Researcher: Dr. Green

Institution: Dalhousie University

DATA COLLECTION

WHAT TYPES OF DATA WILL YOU COLLECT, CREATE, LINK TO, ACQUIRE AND/OR RECORD?

Our team will collect 383+ digital text documents (PDF, Word, plain text), tabular data (Excel), and 900 minutes of mp3 files. We will convert all text documents to plain text format.

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?

Our data will be collected in plain text, CSV, and mp3 form. These formats are industry standard and therefore should allow for easier data re-use, sharing, and 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 GitrKraken for version-control.

DOCUMENTATION AND METADATA

WHAT DOCUMENTATION WILL BE NEEDED FOR THE DATA TO BE READ AND INTERPRETED CORRECTLY IN THE FUTURE?

Our team recommends researchers to keep detailed notes regarding: reflexivity, methodology, methods, and explanations of coding procedures. Versioning software such as GitKracken is has a user friendly GUI while keeping track of any changes to the data and how it occurred. GoogleDrive will also record who has edited what and when. We recommend Dublin Core (DC) for the creation of metadata records due to its broad accessibility.

HOW WILL YOU MAKE SURE THAT DOCUMENTATION IS CREATED OR CAPTURED CONSISTENTLY THROUGHOUT YOUR PROJECT?

GitKraken and GoogleDrive both have automated documentation features to ensure consistent documentation of all researchers' contributions. In addition, researchers will be expected to maintain detailed notes and to provide weekly progress reports to the team.

IF YOU ARE USING A METADATA STANDARD AND/OR TOOLS TO DOCUMENT AND DESCRIBE YOUR DATA, PLEASE LIST HERE.

Metadata will be documented using Dublin Core (DC).

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?

Factoring in for file versioning, backups, and growth over time, our team anticipates requiring at least 200 GB of storage for at least 5 years.

HOW AND WHERE WILL YOUR DATA BE STORED AND BACKED UP DURING YOUR RESEARCH PROJECT?

Our team recommend creating three copies of the data, with one of the three as an external storage source. Dr. Green's practice of carrying a backup on a USB key is good practice, however we recommend switching to an external hard drive located offsite. We recommend storing the data on both Dalhousie's networked drive and on a cloud storage server such as GoogleDrive ($10 per month for 1TB). GoogleDrive is especially useful as a platform for collaborative work so choosing Dalhousie's networked drive as another backup is recommended.

HOW WILL THE RESEARCH TEAM AND OTHER COLLABORATORS ACCESS, MODIFY, AND CONTRIBUTE DATA THROUGHOUT THE PROJECT?

GoogleDrive will allow researchers to collaborate, documents the changes made and who they were made by, and makes this information accessible to the researchers.

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 data will be provided to the funding provider, CIHR (Canadian Institutes of Health Research), for long-term preservation and CIHR will choose who may access it.

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.

Our team will ensure our data is preservation ready by converting the data format to non-proprietary and industry-standard formats: .txt, .mp3, and .csv. We will include all field notes and metadata records. We will also ensure that all interview data has been properly anonymized and codified to prevent identification of study participants. All format conversions will be documented.

SHARING AND REUSE

WHAT DATA WILL YOU BE SHARING AND IN WHAT FORM? (E.G. RAW, PROCESSED, ANALYZED, FINAL).

Due to privacy concerns, we will not share any raw data. However, we will share the processed data once it has been thoroughly analyzed ensuring it is anonymized and codified. Afterwards, we will freely share all analyzed and final data.

CIHR has the following requirements for sharing of data resulting from projects funded by them:

- "ensure that all research papers generated from CIHR funded projects are freely accessible through the Publisher's website or an online repository within 12 months of publication;

- deposit bioinformatics, atomic, and molecular coordinate data into the appropriate public database (e.g. gene sequences deposited in GenBank) immediately upon publication of research results;

- retain original data sets for a minimum of five years (or longer if other policies apply);

- and acknowledge CIHR support by quoting the funding reference number in journal publications". (CIHR, http://www.cihr-irsc.gc.ca/e/32005.html)

HAVE YOU CONSIDERED WHAT TYPE OF END-USER LICENSE TO INCLUDE WITH YOUR DATA?

We recommend using an Open Data Commons license, as endorsed by the CIHR.

"CIHR believes that greater access to research publications and data will promote the ability of researchers in Canada and abroad to use and build on the knowledge needed to address significant health challenges. Open access enables authors to reach a much broader audience, which has the potential to increase the impact of their research. Only when research findings are widely available, enabling open scrutiny, will this evidence be translated into policies, technologies, health-related standards and practices, and new avenues of research that will benefit the health of Canadians and others. From a knowledge translation perspective, this policy will support our desire to expedite awareness of and facilitate the use of research findings by policy makers, health care administrators, clinicians, and the public, by

greatly increasing ease of access to research". (CIHR, http://www.cihr-irsc.gc.ca/e/32005.html).

WHAT STEPS WILL BE TAKEN TO HELP THE RESEARCH COMMUNITY KNOW THAT YOUR DATA EXISTS?

We recommend publishing in an open access journal, such as Sherpa/Romeo, to increase accessibility

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.

While the study is ongoing the data will be stored by the three previously mentioned repositories (DalSpace, Google Drive, and Dr. Green's external hard drive). Our research team of four people will be responsible for transferring, converting, anonymizing and codifying the data. Our team will also produce the accompanying metadata records for each stage of this process. This should be completed within one year; the staff will be trained in all programs and file formats involved.

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?

Of the four researchers on this team our positions are as follows: one will be the principal investigator, and the remainder are co-investigators who will receive the same training and information as the former. This should ensure minimal disruption if there are any unforeseen personnel changes.

WHAT RESOURCES WILL YOU REQUIRE TO IMPLEMENT YOUR DATA MANAGEMENT PLAN? WHAT DO YOU ESTIMATE THE OVERALL COST FOR DATA MANAGEMENT TO BE?

Resources required will include employing a staff of four people for approximately one year, $10 per month for GoogleDrive's 1TB plan, 1 T2B external hard drive. Our team estimates roughly $230,000 in total cost.

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?

All sensitive data will be anonymized and codified by researchers to protect the identities of our participants. The raw data will not be made publically accessible.

IF APPLICABLE, WHAT STRATEGIES WILL YOU UNDERTAKE TO ADDRESS SECONDARY USES OF SENSITIVE DATA?

All participants in our research have been/will be required to fill out a standard consent form, which can be viewed here: https://web.stanford.edu/group/ncpi/unspecified/student_assess_toolkit/pdf/sampleinformedconsent.pdf

Chief Writer: Li Liu - B00639704 (Edited by Brian Jenkins - B00782701)

Group Assignment - Case 3

April 10th, 2018

MY PLAN (PORTAGE TEMPLATE)

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.