Data Consultation Reports

For: Elvira Mitraka, Head Librarian

Database Management Systems

INFO: 6540

Prepared by Research Data Consultants:

Tobbi Dyer
Rachel Fry
Laura Jones
Mark Tambal

Attached are the three Data Management Reports you requested.

Thank you for your interest in our services!

Data Consultation Report for Doctor Periwinkle

SECTION 1: Overview of Case

The following Data Consultation Report has been created for Doctor Periwinkle at Dalhousie University. Dr. Periwinkle leads a large research team made up of students, staff, and researchers (Periwinkle, 2018). She and her team have been conducting research, funded by Innovation Canada (http://www.innovation.ca) on marine wildlife by tracking animals using monitoring equipment and different sensors (Periwinkle, 2018). The data collected is used to make predictions about animal populations and migratory patterns, to write research papers, for graduate school level exercises, and for viewing by the general public (Periwinkle, 2018). This report will provide an overview of what data is collected as well as suggested Data Management Plan (DMP) for how data should be processed, analyzed, preserved, accessed, and reused.

Section 2: Existing Data Formats, Types and Sizes

Several types of collected data will be reviewed for this DMP. The data is collected using the following methods:

1. Using remotely-operated marine vehicles (ROMV) (Periwinkle, 2018).
2. By surgically implanting tags in marine animals that have been captured and released (Periwinkle, 2018).
3. Using “static sensor buoys that measure ocean conditions” (Periwinkle, para 2, 2018).
4. Using “communication lines that passively listen for signals from animal tags” (Periwinkle, para 2, 2018).
5. From field notes that list the animals that have been tagged and released, field notes detailing marine animal population estimates, and field notes written by students in Dr. Periwinkle’s classes (Periwinkle, 2018).
6. Allowing scientists to report wildlife they see in the Minas Basin (Periwinkle, 2018).
7. Running simulation models (Periwinkle, 2018).

The data collected is stored in different formats, each specific to the type of sensor used (Periwinkle, 2018). Data is then converted to NetCDF format using a software developed for the project (Periwinkle, 2018). The following chart shows the type of data collected, the size of the data collected and the format.

Table 1: Overview of Dr. Periwinkle's Collected Data

Type of Data	Size of Data	Format of Data
Data from sensors- ROMV, animal tags, static sensor buoys, communication lines	500MB uncompressed NetCDF data	Network Common Data Form (NetCDF), format 4
Wildlife reports by citizen scientists	3GB	Tab-Separated Values (TSV) Files
Field notes	2GB	Paper notes transcribed to TSV Files Darwin Core Metadata
Simulation models - using data from the buoys and collaborator	200GB (uncompressed)	Zipped Comma-Separated Values (CSV) files
Overall	500GB	total TSV files

Section 3: Facilities and Required Equipment

Data is increasing on an ongoing basis, with the ROMV data being collected monthly. With a large collection of existing data, along with the regular addition of new data, a manageable system needs to be put in place.

Currently the data collected by Dr. Periwinkle and her team does not exist in a centralized location, where researchers can use it without having to request it directly. The data that is currently being worked on and updated will be stored in the cloud on Box.com. Box is an online cloud storage system that will allow for data to be accessed, and worked on collaboratively, from any computer. The Box system that will be used is Business Plus which features unlimited storage, 5GB file uploads, no maximum user limit, unlimited external collaborators, and custom metadata templates (Box a, 2018). This plan costs 35 dollars a month, which will amount to 420 dollars spent annually. The Box storage will be password protected so the only individuals able to access this data will be those on the research team. Users who wish to download specific files to their personal drives and USB devices can download files from the cloud and transfer it to those devices.

The Box storage system was chosen because it provided the most online storage per month for a reasonable cost. We investigated Google Cloud Storage as an option, however the models are pay as you go per GB loaded onto the cloud (Google Cloud, 2018). Dr. Periwinkle's team is working with copious quantities of data which will be updated with each project. Choosing a cloud storage plan with a set price allows for the team to not have to deal with unforeseen budget costs.

The Box cloud storage system is based out of the United States, but none of Dr. Periwinkle's data is of a personal nature, so it was decided that storing it on cloud outside of Canada would be fine.

The data that is not currently active will be stored on a backup server. We advise the team to purchase the Dell Synology Disk Station DS2415+ - NAS server. This server supports up to 144TB of data, features high speed data transfer and encryption capabilities, and has 3.0 and eSATA USB ports which allow for quick file transfer (Dell 2018). The data currently stored on external hard drives and USBs will be easily moved to this backup server. The cost of this server is 1,899.99 dollars (Dell, 2018) and it is a one-time investment. The server will be in the research team's locked office, so that the only people who can make changes to its contents are those who have access to the office.

This server is suggested because it is reasonably priced and will store large amounts of data. A larger server is not necessary because data will be removed from the server when it becomes non-relevant, and so only archived data will be stored on the server. As Dr. Periwinkle currently has data stored in formats such as USBs and DVD’s this data can be burned onto a USB drive and backed up to the server.

The paper data, in the form of the original field notes, will be digitized and uploaded to the Box Cloud and backed up on the backup server.

Section 4: Data Management Practices

Data Analysis: software used

The data to be analyzed is in both a structured and an unstructured form. The software we have chosen to analyze the data is pandas, powered by python, which will be installed on the computers used by the research team. Pandas can handle several types of data but is of use in this scenario for its ability to deal with tabular data, such as CSV and TSV files (Pandas, 2017).

Pandas by Python was chosen as the primary data analysis software because it is free, well renowned, and used by many researchers around the world. Pandas can work with different types of programming languages including Excel and SQL (Pandas, 2017).

Pandas is fast and has been cited as having the potential to become “the most powerful and flexible open source data analysis/ manipulation tool in any language” (Pandas, para 2, 2017). Pandas can read unordered and ordered data, matrix data, observational data, statistical data and more (Pandas, 2017). Pandas is capable of many functions but specifically important to Dr. Periwinkle’s needs are its abilities to find missing data, manipulate data, group data into sets, index data with metadata, reshape data, and merge data sets together (Pandas, 2017). Dr. Periwinkle and her team are interested in using data from other researchers to further their own research. In pandas, data in other formats can be uploaded and merged with the research team’s data and will allow the different data sets to be analyzed side by side.

The remaining digital data is in the NetCDF format. The program used to analyze NetCDF data, NetCDF by Unidata, will also be installed on the computers used by the research team. There is a program add on called netcdf4-python, which is a “python interface to the netCDF library” (Unidata a, para 1, 2018). Netcdf4-pyhton will allow for seamless conversion of NetCDF files so they can be used in pandas. The NetCDF software is free and available on Unidata.ucar.edu (Unidata b, 2018).

The NetCDF program was chosen because it is the only program designed specifically for reading this type of data. NetCDF is an interface specifically for scientific data (Unidata b, 2018). The NetCDF program is compatible with python so the data in these formats can be manipulated in pandas with the python add on. We urge Dr. Periwinkle to update her version 4.0 to version 4.6.1. While the updates primarily concern the maintenance of the software, it is important that her software is kept up to date. This will help to ensure longevity of the data and maintain compatibility with other users of open data (News@unidata, 2018).

Python is free and can be downloaded from the Python software foundation website (Python, 2018). Since python/pandas is difficult to learn, a data analyst will be hired to train the members of the research team. The members of the research team will attend a month-long coding boot camp to learn pandas and python. The cost to send each team member to the session is 5,000 dollars Canadian.

Preserving data: Metadata strategies, Data value after use

The digitized field notes will be kept for five years in their paper form, in an archival room facility within the research facility. The original notes will be destroyed after five years.

Box business plus comes with a metadata feature (Box b, 2018). The metadata feature allows for data to be tagged with the appropriate subject heading, so it can be easily accessed (Box b, 2018).

Dr. Periwinkle will continue to use Darwin Core to create metadata in keeping with the standards set by the Ocean Biogeographic Information System (OBIS, 2018), an organization where she would like to share all of her data. In order for this data to be searchable and retrievable, standardized naming protocols should be adhered to, and it is important that any data provided by students and/or citizen scientists maintains this naming standard. Darwin Core covers metadata regarding the class taxon (name), identification, occurrence, location, event and material sample. (If students are unsure of the details, they should refer to the reference provided at the end of this document).

Dr. Periwinkle has indicated that she only maintains complete data, data that has all fields completed, and we urge her to continue this practise for the integrity of the results. Also, she will need to ensure that she includes the terms and descriptions for organism, feature, depth, environmental conditions and, we would suggest, researcher's name. These tags will help to link the data and make it more accessible and, therefore, more useful.

Coding will be done by the data analysists to ensure the data in pandas has metadata attached to it. Pandas has the ability to index data which will allow Dr. Periwinkle to organize data sets (Pandas, 2017).

Dr. Periwinkle wants to ensure that data that is analyzed can be accessed by the public, and other researchers who are interested in using the data. In addition to sharing data on OBIS, analyzed data and raw data will be shared on The Open Science Framework for free use. The data will be uploaded onto the web with its corresponding metadata.

The data will remain publicly available for ten years online on the Open Science Framework and OBIS, and then it will be digitally archived. This data will then be available upon request from the archives. The data will be archived after ten years to ensure the data being accessed by the public and researchers is still relevant.

Data will be moved from the Box cloud to the backup server once a project is completed. Data will be disposed if it is disproved by new research. While Dr. Periwinkle hopes that data submitted through her website by scientists will be used, if it is exposed to a broader audience, it will be deleted from the open science framework and the backup server if it is not used for five years.

Section 5: Sharing and Licensing Data

Giving access to data: Owner of data, rights to data

The data is owned by Dr. Periwinkle. Currently Dr. Periwinkle has been sharing data with individuals as they request it, but she would like the data to be accessible to researchers and students without them having to contact her. The Open Science Framework is a cloud-based storage system created for sharing scientific research. Dr. Periwinkle and the team will publish reports, field notes that have been digitized and raw data files on the cloud for scientists to analyze. If a researcher whose first language is not English would like access to the data, she has arranged to hire a translator to ensure the accuracy of the data is kept intact. Putting information on the Open Science Framework and OBIS will allow for data that is not currently being used, such as that submitted to her website, to be shared with a wider audience. The access to the data before this plan was limited, so people did not know it existed.

Putting data on open source websites allows for it to reach a wider audience and have a larger impact on oceans' research.

Canada does not currently have a platform for sharing integrated Oceans data the way the United States (IOOS) and Europe (EMODnet) do. A 2016 report published by the Marine Environmental Observation Prediction and Response Network (MEOPAR) overviews the reasons Canada needs an integrated Oceans data sharing platform. Dr. Periwinkle is encouraged to advocate for the creation of this network as it would encourage the sharing of open data, ensure researchers receive credit for their data produced, allow for a streamlined system of ideas based on a geographical region, benefit stakeholders interested in the data, and move towards sustainable Oceans' data management in Canada (Wilson, Smit, Wallace, 2016).

While students have access to Dr. Periwinkle's data, we advise her to ensure that they are aware that the data actually belongs to her and that they can use it under the same licensing agreements as any other users. We will prepare a form outlining this policy them to sign.

Re-using data: Applicable licenses, data reuse

The license that applies to the data on the Open Science Framework is the Creative Commons Attribution 4.0 International (CC BY 4.0) license (Creative Commons, 2018). This license allows individuals to copy and redistribute the material (Creative Commons, 2018) and allows individuals to build upon the data, for non and commercial purposes (Creative Commons, 2018). Individuals using materials protected by this license must give appropriate credit to the owner, and if they do not, they must ask for permission to use it without credit being given (Creative Commons, 2018). Requiring individuals to ask permission to publish data without acknowledgements ensures the original report cannot be changed and re-published with changes, resulting in the preservation of the original research. This license is being used because Dr. Periwinkle wants to ensure the data is open to as many people as possible. This license is international, so researchers around the world have the rights to use the data and do not have to worry about it only being licensed in Canada.

This will ensure that national and international ocean's research teams will be able to use the data to further their own research and will also allow for individual research groups and students to use the data for their studies.

Section 6: Annual Data Processing / Analysis Cost Summary

Box Cloud Storage: $420 CAD

Python: Free

Dell Synology Disk Station DS2415+ - NAS server: $1,899.99 CAD

Python Bootcamp: $5,000 CAD per person = $15,000 - $20,000 CAD

Data Translation Cost: $200 - $500 per project (rough estimate)

Total Baseline Cost: $17519.99 CAD - $22819.99

Plus an addition $5,000 CAD per additional team members (over 4)

We sincerely hope you consider our proposal,

Tobbi Dyer
Rachel Fry
Laura Jones
Mark Tambal

References

Box. (2018a). *Choose a plan that's right for your business*. Retrieved from https://www.box.com/pricing

Box. (2018b). *Box business plus***.** Retrieved from https://www.box.com/pricing/biz-plus

Creative Commons. (2018). *Attribution 4.0 international (CC BY 4.0)***.** Retrieved from https://creativecommons.org/licenses/by/4.0/

Dell. (2018). *Synology disk station DS2415+ - NAS server - 0 GB (DS2415+).* Retrieved from http://www.dell.com/en-ca/shop/accessories/apd/a8343981?ref=p13n_ena_pdp_vv&c=ca&cs=cadhs1&l=en&s=dhs

Google Cloud. (2018). Retrieved from https://cloud.google.com/storage/

News@unidata. (2018, March). *NetCDF 4.6.1*. Retrieved from https://www.unidata.ucar.edu/blogs/news/entry/netcdf-4-6-1

OBIS Ocean Biogeographic Information System. (2018). *Darwin Core.* Retrieved from http://www.iobis.org/manual/darwincore/

Pandas. (2017). *Pandas: Powerful python data analysis toolkit*. Retrieved from https://pandas.pydata.org/pandas-docs/stable/

Periwinkle. (2018). *Case 1_ professor periwinkle*. Retrieved from https://dal.brightspace.com/d2l/le/content/61624/viewContent/942811/View

Python. (2018). *Python*. Retrieved from https://www.python.org/

Unidata a. (2018). *netcdf4 module*. Retrieved from http://unidata.github.io/netcdf4-python/

Unidata b. (2018). *Network common data form (NetCDF)***.** Retrieved from https://www.unidata.ucar.edu/software/netcdf/

Wilson, L., Smit, M., & Wallace, D. W. R. (2016). Towards a unified vision for ocean data management in canada: Results of an expert forum. (). *Halifax Nova Scotia: Marine Environmental Observation Prediction and Response Network (MEOPAR)*. Retrieved from http://meopar.ca/uploads/ODM_workshop_report.pdf

Data Consultation Report for Doctor Green

SECTION 1: Overview of Case

The esteemed Dr. Green has requested that our team create a data management plan (DMP) that will both protect the identity of his research participants, and help him organize his research data. Dr. Green's research, which is funded by CIHR (Canadian Institutes of Health Research) requires that a data management plan is created (CIHR, 2016), and this DMP has been made to comply with the standards outlined by the funding agency in *CIHR Best Practices for Protecting Privacy in Health Research (September 2005).*

Dr. Green is currently working on a project alongside two masters' students, which requires complete confidentiality and the potential risk of revealing participant information is high (Green, 2018). The project we have been asked to review focuses on the hospital as a working environment, and looks at how interdisciplinary primary care teams work together (Green, 2018). While this particular project of Dr. Green's will be the focus of our data management plan, we feel that this plan could be applicable to Dr. Green's other research.

SECTION 2: Existing Data Formats, Types and Sizes

Table 1: Current Data Management Tools in Use

Data Management Tool Currently in Use	**How it is being used**
Zotero	Document sharing
Dropbox	Storing audio files
Google Docs	Hosting interview transcripts

USB key	Master copy/backup
Physical copies	Printed spreadsheets from 2002

Table 2: Current data in use & space requirements

Type of Data	Current Format of Data	Approximate Size of Data
Textual / content analysis *383 digital documents that contain information about the medical teams, outcomes, and practices* *15 interviews transcribed in Word*	Mix of Word, PDF, plain text, and other documents Associated quantitative data is stored in an Excel spreadsheet	If the average text file is approximately 5kb, a Microsoft Office document is 250 kb (Microsoft, 2010), and the average 20 page pdf is 5000 KB (GreenNet, 2016), then the mean file size is 1751 KB. If there are approximately 400 files, than the textual/content analysis should account for around 700,400 kb or **0.7 gigabytes**.
Interview Audio files Data *15 interviews total (presently)*	60 minute MP3 audio files	128kbps each X3,600 second (1 hour)= 460,800kb X15 videos = 6,912,000 kb 1 Mb = 1,024 kb 1 Gb = 1,024 Mb = 6.59 Gb 1 GB = 8 Gb = 0.82 GB

Existing open data about healthcare	No number provided. .tsv files	If there are 100 open data spreadsheets and an office document is approximately 5kb, then there are 500 kbs or **0.0005 gigabytes** of data.
	Approximate size of dataset currently in use by Dr. Green:	**1.5205 GB of data**

SECTION 3: Facilities and Required Equipment

Dr. Green has expressed the importance of protecting the identity of his research participants (Green, 2018), while appreciating the ease and convenience of online storage (Brightspace Discussion, 2018). To accommodate his needs we suggest the following data storage techniques:

Table 3: Data facilities for specific data types

Storage	Type of Data	Example Stored Data
Encrypted Cloud storage on a Canadian Server	Anonymized data	Anonymized transcripts of interviews Anonymized digitized documents Additional open data information Any other helpful digital documents that do not contain identifying information
Encrypted data uploaded to Dalhousie's NAS Drive	For any data that can be traced back to research participants	Audio files of original interviews Digitized originals

Converting Anonymized Data to Cloud Storage

The methods we have selected will ensure both ease of use and the protection of research participant information. We suggest removing all existing data from Google Docs and Dropbox. While we understand that these sites are convenient, they are not necessarily secure (Paul, 2014; U of Vic, 2017). On such websites, data is stored on servers located around the world, including the United States, where certain laws can allow Government agencies to access even the most sensitive of information (Paul, 2014).

In addition, Google Docs are scanned to ensure that they meet the community standards that users agree to when they sign up (Titcomb, 2017). There have been documented instances where researchers have had their work scanned, deleted, and then returned only after a complaint was filed (Titcomb, 2017). This highlights the fact that the content of your files is not blocked from Google itself and that they can be read in their unencrypted form. There is also the risk that your data can be hacked by other users through your Google account (Smith, 2014).

We understand that Dr. Green is currently using Zotero (Green, 2018), and while Zotero is not cloud storage, we would still recommend removing all personal files from this platform. Zotero is an excellent reference and citation manager for bibliographic information and when downloaded onto a desktop can be used to confidentially keep track of files (Zotero.org, n.d.). However, in order to use Zotero as a file sharing platform (as Dr. Green currently does) some information is stored on Zotero's servers which are not encrypted and are located in the United States (Zotero Forums, 2018). Therefore, research data created by Dr. Green should not be stored there.

Nova Scotia falls under the Personal Information Protection and Electronic Documents Act (PIPEDA) which requires that any personal data associated with a government institution be stored in Canada (Server Cloud Canada, 2017). This act pertains to personal information, such as that collected through Dr. Green's studies. Researchers are required by law to protect this personal data and ensure that it is stored under the equivalent to Canadian privacy regulations. For this reason we suggest switching to a cloud storage platform that stores data in Canada.

Cloud Suggestion for Anonymized Data

Given that servers fall under the laws of the countries they are located in, it is highly advisable that Dr. Green store his data in Canada (Server Cloud Canada, 2018). We suggest migrating all anonymized data to sync.com, a cloud service which offers special services to those creating Canadian data on Canadian soil. Sync.com allows one to work online and store documents in the cloud for safety and easy retrieval (sync.com, 2015). Having a central repository for anonymized data will benefit Dr. Green as it will help him share files and collaborate with ease if his project grows as anticipated (Brightspace, 2018). In addition, this proposed storage method will help Dr. Green stay organized and maintain consistent metadata practices throughout the creation and maintenance of his research data.

Files and data stored in sync.com are completely encrypted and cannot be read internally by the company or externally by anyone else, unless you provide a password (sync.com, 2018). This is a direct contrast to major cloud competitors who obtain permission to view your files upon registration. Sync.com uses a two-pronged system to encrypt data. Each file is encrypted with 2048 bit RSA (an algorithm) and 256 bit Advanced Encryption Standard (AES), and then Transport Layer Security (TLS) is applied as additional protection for data transfers. This will ensure the optimal protection and privacy for your documents (sync.com, 2018).

To minimize the risk of lost data or files, sync.com has a robust version control system (sync.com, 2018). You can easily view the file revision history and select from any previous version of your document that was edited online, and restore deleted documents. This is especially important when working as a team so that you can ensure that you can always access previous versions if a mistake or corruption is detected (sync.com, 2018).

Storage of Highly Sensitive Data

Highly sensitive data that includes any identifiable information (including audio files of participant voices) should not be stored in third-party cloud storage (Concordia University, n.d.). And while Sync.com offers a great deal of privacy protection as described above, it is possible that policy changes or other commercial issues could affect your data (Concordia University, n.d.). For this reason, all original and identifying datasets should be stored on Dalhousie's NAS Drive.

While Dr. Green has expressed some reservations about using Dal's existing data management infrastructure (Brightspace, 2018), we feel that this option is best because it is the most secure option as the servers are owned by Dal. Furthermore, the university recommends that faculty and students conducting research within the institution store any and all identifiable data there (DRE & ITS, 2017). This data management standard implemented by Dalhousie is in compliance with The TCPS2 (Tri-Council Policy Statement 2014) which is a requirement of Dr. Green's funding agency.

Manual Encryption

We advise that all confidential files, and all files uploaded to a third-party cloud service, be encrypted with software that uses Advanced Encryption Standard (AES) algorithms. We recommend Folder Lock which recently received an “Editor’s Choice’ and “Excellent” rating in PC Magazine (Rubenking, 2016). Dr. Green will be able to secure files, folders, and portable devices through encryption. In addition, he can protect emailed files by creating password protected zips if required. The MP3 files can also be zipped and locked before being stored and backed up.

These encryptions can only be accessed through a password or, as a backup, through the software’s serial number (Rubenking, 2016). Rubenking cautions, however, that with the company having a record of your serial number, this could potentially be accessed by governments or other law enforcement organizations. We advise Dr. Green to allow us to disable this feature (using the serial number as a backup) in the settings when the software is installed.

SECTION 4: Data Management Practices

Creating Backups of All Files

Backing up data simply means ensuring that there is another copy in case the first is accidentally destroyed, deleted or corrupted. We have created a solid system of backup that will ensure Dr. Green never loses any important research data:

1. All identifiable data that is stored in Dalhousie’s NAS Drive be backed up onto an encrypted file of Dr. Green’s personal desktop computer located in a locked office. This can be initialized as an automatic back up.
2. All anonymized data currently stored in the sync.com cloud also be backed up onto Dr. Green’s encrypted personal desktop computer located in a locked office. This can be initialized as an automatic back up.
3. A manual back-up on an external hard drive kept in a fireproof locked safe in Dr. Green’s locked office which is bolted down so it cannot be removed. This manual back-up should be performed twice weekly.

In addition, we suggest checking on a bi-weekly basis, when manual back-up is conducted, that files are readable and properly saved (University of Ottawa Library, n.d.). This bi-weekly schedule would also be an excellent time to ensure proper metadata and archiving practices are in use as described below.

Converting Data to Interoperable Formats

We suggest converting current data to as few file formats as possible. Converting all files to compatible formats will allow for greater interoperability and will allow Dr. Green to search through files using the “ctrl+f” function. It will also simplify the process of data collection and enable more consistent metadata practices.

All existing quantitative data should be converted to Excel files, and all qualitative data should be converted to Word files whenever possible. This includes all existing text files and pdfs. Existing text files can be converted into Word documents by simply pasting the files into the word files and saving them using the naming strategy described below. In order to convert pdfs, we suggest using the following site: http://pdf2doc.com/

In addition, files that are not regularly used (such as the interview MP3s once they have been transcribed and the original copies of anonymized data) can be zipped and archived.

Currently, Dr. Green has a binder containing Excel documents from 2002. We suggest that Dr. Green consider whether it is worth investing the time to scan, and upload these files so that they are digital, or possibly find this information online and already digitized. If they are not accessible, or you deem their digitization unnecessary, we suggest you consider destroying the files. Keeping unnecessary research data will only cloud the research process, and this new streamlined data management plan which has been written to comply with the rules outlined by Dr. Green’s funding agency will help Dr. Green create better, more useful research data.

File Naming, Metadata and Organizing Strategy

All staff working with Dr. Green will need to have thorough training on the creation of metadata associated with research files. Metadata may seem like an unnecessary step, however it allows for easier retrieval, better organization which ultimately saves times, and ensures the protection of the data. It also allows you to more easily manage data as time progresses.

A file naming strategy should be created and consistently used to ensure files are easy to find. We suggest visiting the following website: Stanford University's “Best practices for file naming” which can be found here. It suggests creating a date format (i.e. YYMMDD), a naming template that includes information about who saved the file, what project the file is for, the type of the file, etc. For example, for interview transcripts, the following naming structure could be used:

[Interviewsubjectcode]011_InterviewerInitials_YYMMDD.doc

This standard, or a similar one should be used when all new files are created, as well as the conversion of existing files to a unified format.

Additional metadata can actually be added right into Word and Excel files. This little known trick can save you hours of searching through documents, and creating metadata standards for the research team. Instructions for doing so can be found here. Adding metadata to Word and Excel is a great way to add extra information to files, and it is searchable without opening the file. For example, the author of a particular document can be added as a metadata feature, as can the date it was created, and you can add custom properties such as “Comments” or “Keywords”.

Files should be organized in a well thought out and hierarchical structure. This will help to keep track of data, and creating the structure ahead of time will save any confusion later and help maintain consistency. Again, maintenance will need to be conducted regularly in order to ensure metadata is properly recorded. We suggest that when bi-weekly back-ups are performed, you check to ensure metadata procedures are being followed. In addition to training, we suggest creating a file that all researchers have access to which clearly explains all expected metadata procedures.

Data Disposal

Dr. Green should dispose of data in accordance with the policies and guidelines laid out in the terms of his funding. However, if no terms were ever made it is reasonable to assume that data would be kept for five years after the publication of the study. If he does need to dispose of data that contains private information, it cannot simply be deleted (UK Data Service, 2018). To completely delete a file, you must overwrite it using a software such as Eraser (UK Data Service, 2018) or Folder Lock, as described above. This will ensure that private information of research participants will never been seen by anyone who should not have access to it.

SECTION 5: Sharing and Licensing Data

Sharing Data

Once the Data Management Plan has been finalized, we suggest updating your confidentiality form for both future research participants and researchers. This will ensure that all those involved in the project understand the new, exciting DMP which will help protect your research participants and make this data easier to work with.

When a new confidentiality form has been signed, Dr. Green can grant access to information as he pleases. He can easily protect the identifiable data stored on Dalhousie's NAS Drive by not providing a password to any or all members of his research team. This would allow researchers to explore anonymized data stored on Sync.com, create and update metadata, while still allowing Dr. Green to maintain control over the original files that contain confidential information. Within Sync.com, various permissions can be granted and certain researchers could be given access to only particular files (sync.com, 2018).

As Dr. Green indicated, some data requires translation to the official language of the collaborator's location (Brightspace, 2018), which will allow them appropriate access. A data translator will be hired for each language required. This translator will have to sign and adhere to a confidentiality agreement so that personal data is not shared either inadvertently or intentionally.

Licensing Information

Currently Dr. Green has requested that his research data remain private, and we understand that this is a legitimate concern. Faculty who conduct research at Dalhousie do maintain the right to their research as specified under Article 23 of the Dalhousie Faculty Association collective agreement (Dalhousie Copyright Office, n.d.). However, as Dr. Green's project progresses, we suggest that sharing portions of his anonymized data as part of a Creative Commons License would be very beneficial to the public. This license would allow others to view and reuse, but credit will always be given to Dr. Green as the original creator.

We feel strongly that aspects of this data could help other hospitals create better, more productive teams. More and more, funding agencies and research institutions are pushing those who gather important information to open their data sets up to the public (Borgman, 2012). In fact, the need for researchers to show their datasets to the general public has been referred to as "urgent" (Borgman, 2012, p. 1060) because it encourages transparency and allows researchers to connect with one another to build stronger research communities. Not all of Dr. Green's research data would need to be included in this license. Dr. Green would be able to release only certain aspects of data that he felt would benefit others. When a licensing agreement is made, a contract can be drawn up to restrict certain aspects of the data from reaching the public (Government of Canada, 2010). This is called an Information Sharing

Agreement (Government of Canada, 2010)

Furthermore, as Dr. Green publishes articles he would be able to copyright his journal articles published through the Scholarly Publishing & Academic Resources Coalition (SPARC) Author Addendum (Dalhousie Copyright Office, n.d.). And register books, and other published material under Canadian Copyright Law. See the guide to copyright law here.

SECTION 6: Costs

Dr. Green does not require a great deal of storage space for his research data, which reduces the cost of associated data management tools. In **Table 2,** we indicate that currently Dr. Green is only using about 1 GB of data. We suggest that Dr. Green invest in higher storage levels than might be immediately necessary. This will ensure that he never has to worry about reaching capacity.

Table 4: Predicted costs (both monthly costs and predicted expenditures)

Purchase Item	**Purpose**	**Cost**
Sync.com space	For sharing anonymized data between researchers	$5.00 (per month for 1 TB of storage)
Folder Lock	For encrypting data	$39.99 (one time purchase)
External Hard-drive	Data backup	$59.98 (one time purchase)
Fireproof Safe that can be bolted down	For storage of external hard drive or any other sensitive documents	$168 (one time purchase)
	One Time Purchase Total:	$267.97
	Monthly cost Total:	$5.00

There will also be costs incurred for data translation. This cost could range from $500 - $10,000 depending on the number of languages and the complexity and amount of the data.

Summary, Longevity & Future Use

Our DMP does not complicate Dr. Green's current practices, it simplifies them. With fewer repositories for data, a solid plan to backup data so it will never get lost, and simple and new metadata strategies, the research data for this project will be safely stored, organized and safe. Dr. Green can also relax knowing that private data is protected, and that when ready, anonymized data can be shared with the public.

We sincerely hope you consider our proposal,

Tobbi Dyer
Rachel Fry
Laura Jones
Mark Tambal

References

Borgman, C. (2012). The conundrum of sharing research data. *Journal of the American Society for Information Science and Technology, 63*(6), 1059-1078. doi 10.1002/asi.22634

Brightspace Discussion. (2018). Retrieved from INFO6540 Brightspace Discussion

Board on 8 April 2018.

CIHR. (2005). *CIHR Best Practices for Protecting Privacy in Health Research*

(September 2005). Retreived from http://www.science.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html?OpenDocument

CIHR. (2016). *Tri-Agency Statement of Principles on Digital Data Management*. Retreived from http://www.science.gc.ca/eic/site/063.nsf/eng/h_83F7624E.html?OpenDocument

Concordia University. (n.d.). *Data storage and file formats.* Retrieved from https://library.concordia.ca/help/data/data-storage.php

Creative Commons. (2018). *Attribution 4.0 international (CC BY 4.0)*. Retrieved from https://creativecommons.org/licenses/by/4.0/

Dalhousie Copyright Office. (n.d.). *Retaining Your Copyright*. Retreived from https://libraries.dal.ca/services/copyright-office/for-faculty/retaining-copyright.html

Dalhousie Research Ethics and Information Technology Services (DRE & ITS).(2017). *Protecting electronically stored personally identifiable research data*. Retreived from https://cdn.dal.ca/content/dam/dalhousie/pdf/research-services/REB/Protecting_Electronically_Stored_Personally_Identifiable_Research_Data.pdf

Government of Canada. (2010). *Guidance on preparing information sharing agreements involving personal information.* Retrieved from https://www.canada.ca/en/treasury-board-secretariat/services/access-information-privacy/privacy/guidance-preparing-information-sharing-agreements-involving-personal-information.html

Green. (2018). *Case 2_professor_green.* Retrieved from https://dal.brightspace.com/d2l/le/content/61624/viewContent/942812/View

GreenNet. (2016). *Understanding file sizes*. Retrieved from https://www.greennet.org.uk/support/understanding-file-sizes

Microsoft. (2010). *Performance and capacity planning (FAST Search Server 2010 for SharePoint)*. Retrieved from https://technet.microsoft.com/en-us/library/gg604780.aspx

NewSoftwares.net. (2018). *Folder Lock*. Retrieved from http://www.newsoftwares.net/folderlock/

Paul, D.A. (2014). *Saving your files: cloud or network?* Retrieved from http://www.canadianlawyermag.com/article/saving-your-files-cloud-or-network-2455/

Rubenking, N. (2017). The best encryption software of 2018. *PCMag Digital Edition*, Retrieved from https://www.pcmag.com/article/347066/the-best-encryption-software-of-2016

Rubenking, N. (2016). Folder lock. *PCMag Digital Edition.* Retrieved from

https://www.pcmag.com/review/347453/folder-lock

Server Cloud Canada. (2018). *When your data must stay in Canada*. Retrieved from https://www.servercloudcanada.com/2017/10/privacy-law-canada/

Smith, M. (2014). How secure are your documents in google drive? *MakeUseOf.com.* Retrieved from http://www.makeuseof.com/tag/waze-hands-free-navigation-drivers/

Stanford Libraries. (n.d.). *Best practices for file naming.* Retreived from https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

sync.com. (2015). *Privacy white paper*. Retrieved from https://www.sync.com/pdf/sync-privacy.pdf

sync.com (2018). *What is file version history?* Retrieved from https://www.sync.com/help/what-is-file-version-history/

Titcomb, J. (2017) Why Google is reading your docs. *Technology Intelligence*. Retrieved from https://www.telegraph.co.uk/technology/2017/11/01/google-reading-docs/

UK Data Service. (2018). *Data disposal*. Retreived from https://www.ukdataservice.ac.uk/manage-data/store/disposal

University of Ottawa Library. (n.d.). *Storage of active data, backup and security*. Retrieved from https://biblio.uottawa.ca/en/services/faculty/research-data-management/storage-active-data-backup-and-security

University of Victoria. (2017). *Data security & cloud storage*. Retrieved from https://www.uvic.ca/systems/support/informationsecurity/datasecurity/datasec-cloudstorage.php

Zotero. (n.d.). *Zotero: The Basics*. Retrieved from https://www.zotero.org/support/quick_start_guide

Zotero Forums. (2016). *Security Questions*. Retreived from https://forums.zotero.org/discussion/47837/security-questions

Data Consultation Report for Doctor Pinkerton

SECTION 1: Overview of Case

Our group of research data consultants has been formally requested by Dr. Pinkerton to prepare a Data Management Consultation Report towards a Data Management Plan (DMP). The purpose of this report is to make recommendations for the primary repository for data along with a cloud storage system in order to manage her large collection of Excel files. The goal of the DMP would be to enable most of Dr. Pinkerton's data to be shared on an open source platform while protecting sensitive data as required. Our plan will follow the data lifecycle and will encompass the creation of data, the processing and analyzing of data, preservation of data, providing access to data, backing up data, and reusing data. Neil Gaiman, in a break from his writing career, will work with us to build and manage the folder structure.

SECTION 2: Existing Data Formats, Types and Sizes

Table 1: Overview of Dr. Pinkerton’s collected data:

Datasets	Data Type	Current Format
Systematic literature review	Quantitative textual analysis	.xlsx
Job descriptions - entry level positions	Unspecified	.xlsx
Student performance data	Unspecified	.xlsx
Downloaded open source data	Stored for future use	.xlsx

The number of rows in Dr. Pinkerton's spreadsheets range from small to very large.

1. Smallest - 1 row (Pinkerton, 2018).
2. Median - 1000 rows and 10 columns (Pinkerton, 2018).
3. Largest - 750,000 rows (Pinkerton, 2018).

In total Dr. Pinkerton has 17,384 Excel spreadsheets, and these numbers will increase as she downloads more data sets (Pinkerton, 2018). The oldest spreadsheet is nine years old and the newest spreadsheets are from 2018 (Pinkerton, 2018).

It is difficult to estimate the amount of space Dr. Pinkerton's files take up because the exact number of spreadsheets she has is unknown, but the average size of an excel document is 14KB (Harkins, 2007).

Excel 2016 has 2gb of storage space (Microsoft, 2018) and the number of workbooks able to be created, relies on how much space is available on the computer (Microsoft, 2018).

SECTION 3: Facilities and Required Equipment

Processing and Analyzing data

Currently, the professor's collected data is stored on her laptop. Dr. Pinkerton has enlisted the help of a postdoctoral fellow, Neil Gaiman, to build a folder on the University cloud storage system to manage her files (Pinkerton, 2018). The cloud storage system Dalhousie University is using is Dataverse (Dalhousie libraries, 2018a).

Dataverse is one of the institutional repositories used by Dalhousie University (Dalhousie Libraries, 2018a). Dalhousie University faculty can request an account by contacting the research data management team (Dalhousie libraries , 2018b). Storage capacity and upload size, varies by format. Dr. Pinkerton can refer to Dalhousie support staff, if there are issues with data uploading (IQSS, 2018b).

Dataverse has the following features:

- Supports SPSS, STATA, R, Excel and CSV files (IQSS, 2018a)
- Data Citation (IQSS, 2018e)
- Multiple publishing workflows (IQSS, 2018e)
- Account Notifications (IQSS, 2018e)
- Three levels of metadata (IQSS, 2018e)
- Searching capabilities (IQSS, 2018e)
- Ability to create restricted and non-restricted files (IQSS, 2018e)
- Ability to customize the datasets (IQSS, 2018e)
- Data analysis capabilities (IQSS, 2018e)
- Ability to display geospatial data (IQSS, 2018e)

SECTION 4: Data Management Practices

Providing Access to Data

Dr. Pinkerton expressed that she is open to sharing her data upon request (Brightspace, 2018). We appreciate this and will design the DMP to ensure that she can easily share her data with others. While Dr. Pinkerton can restrict access to only herself or any combination of colleagues, we suggest that she make existing Open Data sets available for public access.

Student data, however should not be uploaded to Dataverse unless sufficiently anonymised. Depending on the original agreement between Dr. Pinkerton and the students who offered their data, Dr. Pinkerton may also want to restrict access to files containing student data. This can be done easily in Dataverse by creating Terms of Access (IQSS, 2018b).

Preserving Data

The purpose of creating metadata is to maximize the findability and usability of relevant information (University of Alberta libraries, n.d.). To facilitate the creation of useful metadata, Dataverse has several methods in place.

For example, Dataverse encourages users to upload their data using "standard-compliant metadata" (IQSS, 2018c). While this might not seem immediately necessary, using standard metadata schemes, particularly when working with open data, will ensure a universal understanding and interoperability (IQSS, 2018c). The suggested terms are clearly described in the Appendix of Dataverse' User Guide which can be found here.

Additional metadata can be added using "Tags" (IQSS, 2018b). These Tags can be assigned to Dr. Pinkerton's data to help distinguish between data sets once they are uploaded. The tags could represent the theme of the data, or the type of Data. For example, the tags, "systematic literature review" or "student data" could be assigned to associated datasets. Clicking on one tag shows all data linked to it which is an excellent way to manage your own data (IQSS, 2018b), and also find similar data sheets.

CSV files require additional documentation as suggested in the Open Data Handbook. They state that while CSV (comma-separated formats) are excellent for large datasets because they take up little space and ensure a certain level of interoperability, they require additional documentation, or metadata, so that those who open the file can understand it (Open Data Handbook, n.d.)

We suggest creating a text file that would accompany each CSV file that contains information about how the data was collected, why it was collected, and what the structure of the file is. For example, if the file is contained within a table, the rows and columns of the table should be described in the documentation. Her CSV files, generally obtained from open sources, could later be added to an Excel spreadsheet, and this will allow her to find the appropriate file to use.

When storing information in Dataverse, a hierarchical file structure should be followed as should a standard naming template (Stanford Library, n.d.). The naming template can be anything that works with Pinkerton's preferences, however should include standards such as date format and should clearly indicate what should and should not be included, etc. For example:

Nameofdataset_Initialsofuploader_dateuploaded(yymmdd).csv

All of those working with these files should be aware of the metadata procedures and their importance. This is particularly important when working with open data because it helps to ensure that the information maintains its value over time.

In order to ensure metadata integrity, files should be regularly opened to ensure they are readable and properly saved (University of Ottawa Library, n.d.), and we suggest doing this on a weekly basis. We also advise that a .txt file explaining the metadata procedures be included with the data set to ensure likelihood of adherence.

Backing-Up Data

Dr. Pinkerton has amassed 95% of her spreadsheets from outside sources; she has expressed concerns that these sources may not provide access to the data in the future, so she has been storing the data on her local computer (Pinkerton, 2018).

We are proposing that Dr. Pinkerton store backup versions of the data on her central desktop computer. Dataverse has created script that can be scheduled to create an archived copy of a data file as it is updated on the cloud (IQSS,2018d). With the Dataverse script, Dr. Pinkerton will not have to worry about manually uploading the data to an external hard drive herself, and will not lose versions of it.

SECTION 5: Sharing and Licensing Data

Re-using data

Dr. Pinkerton is contacted regularly and asked about data files by fellow researchers who are interested in using these data sets (Pinkerton, 2018). Dr. Pinkerton has been emailing spreadsheets back and forth with people, and has yet to formalize a way for this data to be shared (Pinkerton, 2018).

We are encouraging her to license her data with the MIT license. The MIT license is an open source license which allows individuals who have access to data licensed under it to copy, merge, modify, sublicense, distribute, and sell it (OpenSource, 2018). The data can only be used and modified if the MIT license and copyright date is acknowledged when it is reused (OpenSource, 2018) and will allow her to make the data sets openly available on the Dataverse for public use. This would eliminate the need for people to contact her directly to request data sets, and would free up some of her valuable time. As noted above, all student data can be restricted to protect the anonymity of the individuals.

By making the data open source, more people will be able to use it to further their research. Currently Dr. Pinkerton has many spreadsheets which she has never used, however she is interested in eventually using the spreadsheets to answer research questions and write papers (Pinkerton, 2018). Other researchers may be interested in using the data for similar reasons but if Dr. Pinkerton does not disclose the depth of her collection of spreadsheets, other researchers will not know this data exists. By making the spreadsheets available as open source datasets, the data can then be used for external research which will increase its value. Data in itself has little value if it is not being used, and by making it open source, its reuse can be enabled long term allowing it to be further analyzed and verified.

Under the MIT license, anyone reusing the data will have to acknowledge that it belongs to Dr. Pinkerton. Further, since the data will remain available through open source, and a backup copy will exist on Dr. Pinkerton's computer, the original data will be well preserved. Dr. Pinkerton can rest assured that while the data can be used for other projects, the original research will remain permanently available.

We sincerely hope you consider our proposal,

Tobbi Dyer
Rachel Fry
Laura Jones
Mark Tambal

References

Dalhousie Libraries. (2018a). *About the Dalhousie University Dataverse*. Retrieved from http://dal.ca.libguides.com/rdm/daldataverse

Dalhousie Libraries. (2018b). *Create an account*. Retrieved from http://dal.ca.libguides.com/rdm/getstarted

Harkins, S. (2007). *How to recover from excel workbook bloat*. Retrieved from https://www.techrepublic.com/blog/microsoft-office/how-to-recover-from-excel-workbook-bloat/

Institute for Quantitative Social Science. (2018a). *Supported file formats*. Retrieved from http://guides.dataverse.org/en/latest/user/tabulardataingest/supportedformats.html

Dalhousie Libraries. (2018a). *About the Dalhousie University Dataverse*. Retrieved from http://dal.ca.libguides.com/rdm/daldataverse

Institute for Quantitative Social Science (IQSS). (2018b). Dataset + file management. *Dataverse User Guide*. Retrieved from http://guides.dataverse.org/en/latest/user/dataset-management.html

Institute for Quantitative Social Science (IQSS). (2018c). Appendix: Metadata References. *Dataverse User Guide*. Retrieved from http://guides.dataverse.org/en/latest/user/appendix.html

Institute for Quantitative Social Science. (2018d). *Backups*. Retrieved from http://guides.dataverse.org/en/latest/admin/backups.html?highlight=backup

Institute for Quantitative Social Science. (2018e). *Features*. Retrieved from https://dataverse.org/software-features

Microsoft. (2018). *Excel specifications and limits*. Retrieved from https://support.office.com/en-gb/article/excel-specifications-and-limits-1672b34d-7043-467e-8e27-269d656771c3

Brightspace. (2018, April 4). *General questions*. Retrieved from https://dal.brightspace.com/d2l/le/61624/discussions/threads/232381/View

Open Data Handbook. (n.d.). *File Formats*. Retrieved from http://opendatahandbook.org/guide/en/appendices/file-formats/

Open Source. (2018). *The MIT license*. Retrieved from https://opensource.org/licenses/MIT

Pinkerton (2018). *Case 3 Pinkerton*. Retrieved from https://dal.brightspace.com/d2l/le/content/61624/viewContent/942813/View

Stanford Library. (n.d.). *Best practices for file naming*. Retreived from https://library.stanford.edu/research/data-management-services/data-best-practices/best-practices-file-naming

University of Alberta Library. (n.d.). *Archive my research data*. Retrieved from https://www.library.ualberta.ca/research-support/data-management/preserve

University of Ottawa Library. (n.d.). *Storage of active data, backup and security*. Retrieved from https://biblio.uottawa.ca/en/services/faculty/research-data-management/storage-active-data-backup-and-security