Group Project — Data Management Plan

Colleen Savage, Courtney Vienneau, Nicole Walsh, Randolph White

INFO 6540 – Database Management Systems (XL)

Professor Elvira Mitraka

April 10, 2018

Data Management Consultation Report for Professor Periwinkle

Dr. Periwinkle, along with a large team of researchers, students and staff, participate in the tracking of marine wildlife. In order to efficiently manage and store all data collected, it is important to consider the following data management plan. Dr. Periwinkle's data is collected in a variety of ways. Data is collected digitally by remotely-operated marine vehicles, static sensor buoys and signals from animal tags. All data collected is then converted to NetCDF format for use and storage. The data collected daily is approximated to be 500 mb in uncompressed NetCDF files. Other data collected includes wildlife sighting reports, as well as field notes documenting animals captured and tagged, mark-recapture population estimation experiments and observational studies. Additionally, simulation model data is produced and is in a zipped CSV file format. Citizen Scientist data is also download as a TSV file. Currently the data is made available to the general public through oceanviewer.org, however, the data is only available in a visual form. It is also our understanding that a long-term goal of Dr. Periwinkle is to share data outside of the university and make it available for use. Currently, Dropbox is used to make data available to others that do not have access to the university's file sharing program. Those who have access to the data include Dr. Periwinkle, researchers, students, staff and others who have requested access from Dr. Periwinkle by email. Data is currently shared within the team using external hard drives and USB keys. Also, there is a help document on a shared drive that will need to be located and made available for viewing. Dr. Periwinkle's personal collection of data, dating back to 1998, is in a variety of formats (floppy disks, CDs, DVDs etc.) and will need to be converted into a single digital format and stored in a secure location with the

rest of the data. Our team recommends that all data follow a single metadata standard to increase access through queries. Because a significant amount of Dr. Periwinkle's research data includes field notes, observations and other information related to marine life, we recommend that the entire team use the Darwin Core metadata standard. This metadata standard is also used on the IOSBIS.org website. It is our recommendation that all data is made available online, using cloud storage and online collaborative services. This would enable the team to maintain version control, as many cloud computing services allow multiple users to view and edit data as well as save different versions throughout the editing process. Due to an increase of funding, there has been a substantial increase in the amount of data collected. In order to manage the large amounts of data, we recommend subscribing to a secure data repository. A possible option would be DigitalOcean, which not only provides cloud storage but also offers collaborative team service that allows various members of a team manage data. Our team also recommends that all data is stored and accessed in one secure location. DigitalOcean fulfills this requirements by using a multi-factor authentication process for access to data, which would ensure data security for Dr. Periwinkle and the research team. There are also options for backing up and recovering lost data. The cost for the DigitalOcean system is approximately 5 dollars per month, per 250 GB of storage after a two-month free trial. One of the benefits of DigitalOcean is that they provide flexible pricing, in which users can opt to change their choice of storage model, or droplet model, to better suit their needs. In order to accomplish this, all file formats would have to be converted to JSON format. There are a variety of converters available online at little to no cost. However, if converting all files to JSON format is not ideal, we would

recommend that Dr. Periwinkle to consider using BigQuery data storage warehouse. While NetCDF, CSV and TSV files are not readable on the cloud, they can be stored and later downloaded to a local drive for viewing. BigQuery allows teams to access and manage data collaboratively and in real-time. BigQuery is also free for up to 1TB of data analyzed each month and 10GB of data stored. BigQuery also allows teams to communicate and manage data in real-time and provides data encryption and disaster recovery. Both resources employ vital data management practices such as access control, data recovery, and long-term preservation through the archiving of data. Additionally, if Dr. Periwinkle is given the opportunity to engage in a data sharing agreement or publish collected data, an open data license would be ideal.

The cloud storage options suggested in this consultation provide flexibility, which is beneficial for Dr. Periwinkle and her research collaborators as data from ROMV, static sensor buoys, field observations, and animal tracking tags can quickly accumulate. We believe that converting the variety of data formats Dr. Periwinkle and her team have produced over the years into a JSON format will be necessary to ensure long-term preservation. Lastly, the collaborative opportunities available through online communities such as DigitalOcean will provide other researchers with the opportunity to re-use the data.

Data Management Consultation Report for Professor Green

In this project, Professor Green and his team, two Masters students, will be collecting research data regarding teamwork in hospital environments. They have two primary means of data collection. The first is a textual/content analysis of 383 documents, currently stored on Zotero in various formats. Quantitative data regarding the textual analysis is stored in a Microsoft Excel spreadsheet. The second type is audio recordings and transcriptions from interviews with informants; audio recordings will be encoded in MP3 format at 128kbps. Professor Green will also utilize open source datasets regarding healthcare in his research. Currently, Professor Green stores his data in three locations: Zotero, Dropbox, and Google Docs. Our data management team suggests purchasing a secure database for storage. The first option we suggest is that Professor Green utilizes Dalhousie's' institutional repository, which provides a stable URL, access control, and is easy to use. It can handle the different file formats Professor Green uses, and he will be able to directly analyze his data. The second option is purchasing a DropBox Business account for teams. The standard account costs $17.50 a month, and provides 2 TB of storage, which will provide enough storage as Professor Green's research project grows. A DropBox Business account will provide storage, security, access from different locations, version history and file recovery, and advanced sharing permissions that Professor Green can control. The third option is purchasing a G Suite Business account, which has 1TB of cloud storage, administration controls, archive, and is $10 USD a month. Our suggestion is DropBox, as Professor Green is familiar with it, and can control the access to data. The data created by Professor Green's research is highly sensitive, and he does not want this data shared

with anyone. Any data created by his research will be encrypted before being deposited into DropBox, and will have any identifying information anonymized. Original data will not be shared, but notes recorded by Green himself will be made accessible to his Masters students.

To improve the use of Green's data, we will implement a metadata standard. Data created by his project will be tagged in XML using the Data Documentation Initiative format. This is a widely used international standard for describing data from both social and science fields. Metadata will provide context about data for his team since they will not be allowed to access data without Green's approval. Green will hold intellectual property rights for all his research data and information that is directly created by the project. Green is adamant that the data for this experiment is held closely. However publishing his data with a copyright license will ensure that his data is not used by anyone. To protect the privacy of the informants, Professor Green will use an informed consent form. The form states that the researchers will not identify the participant by name in any information resulting from the interview, and that confidentiality will be kept at all times. Data is subject to the standard practice of anonymizing data to ensure this privacy; any identifying information will be limited to Professor Green and will only be kept on the master copy on his USB key. During analysis and storage, only Green and approved members of his research team will be allowed to access sensitive data. After being uploaded to DropBox, any direct identifiers will be removed.

Data obtained from healthcare organizations are currently stored in a Zotero database in various formats. To ensure preservation of data, it is suggested that the document

files and transcripts of interviews be converted into a single file format of PDF. This will not alter the content of data in any form, but will ensure that the documents can be accessed in the future, and that the data will not be altered. PDF provides a high level data encryption and offers a secure way to transfer data over the internet. Green can customize the user access level, setting it so he is the only one who can access the files and add a digital signature to further protect his work. Phone interviews will be recorded as MP3 files. Each audio file is estimated to be 57.6 MB, for a required 864 MB total storage space.

Preservation and archiving Green's data will occur through DropBox and a master copy of all files will be backed up on both an external USB key and a hard-drive. Green's assigned data manager will be tasked with migrating research data to new formats, platforms, and storage as required for the continuation of his research. A master copy of each file will be contained in the USB key and backed up on the repository in a secure zipped folder. The folder containing files for his project will be both encrypted and password protected. The data files from Green's project will be managed, processed, and stored in a secure environment by the researcher himself. As the data managers for this project, we will instruct Green on how to secure his laptop and install a strong firewall system to further protect his data. This laptop can be used for any project meetings or any potential presentations facilitated by Professor Green. Professor Green has assigned us to be the data managers of this project, meaning we will act as stewards for the data throughout the data life cycle. At the conclusion of the current grant, we will give all data management plan documents to the new assigned data manager.

We recommend that once Green makes use of the documents from the health care organizations for textual analysis, the files will be archived and retained for 1 year after this part of the project is completed. Transcriptions of the interviews will be retained by the repository as part of the permanent research data, and the MP3 files can be archived after they are transcribed. We also recommend that Green's trained Masters students go through the binder of printed Excel spreadsheets from a previous study to find any relevant data to the current research. We will digitize any relevant data to the repository.

A private GitHub account will be used in tangent of DropBox to manage version control during the course of Green's research; it will also act as a backup of any data. A standardized naming convention will be employed for different versions of Green's data. They will consist of a determined prefix, root and suffix system. This will ensure that Green and his team know which is the most recent version of their data analysis and allow the team to access previous versions at any time. Separate files will be managed for the different types of data produced; I.e. there will be one folder for documents, interviews, and notes. Costs for creating, maintaining, and storing data will be covered by a grant Professor Green has received from CIHR. Budget resources will go towards paying for a digital repository, DropBox, and ensuring version control for the long term plan of the research project.

Data Management Consultation Report for Professor Chartreuse

This project generates science of science research data that will be gathered by Professor Chartreuse, distributed to his colleagues for collaborative purposes, and to his graduate students at JCU for educational purposes. The data is predominantly from

public domain resources, and will maintain the licensing of PubMed, which is the database Professor Chartreuse deems most valuable to his research. The data will require different levels of accessibility due to their sensitive nature. Professor Chartreuse and his colleagues require full access to the data, and students should be able to download the data without altering the original dataset. The data generated by the research team will be a revisable dataset. Professor Chartreuse is solely responsible for data storage and providing access to other parties. At the moment, Professor Chartreuse stores all of his data in large Excel files, which creates an obstacle as Excel files have a large margin of error within the dataset, and are difficult to use for sharing with his team. Our data management team suggests that the 20GB of previously generated data will be converted from JSON to CSV files, while new data that is produced will be saved as CSV in order for the files to be machine readable. Prior to converting the 20GB of previously accumulated data, we suggest working with the research team to assess what data is needed, what data should be disposed of, and the data that should be archived. We will also work to anonymize any data to ensure they do not contain any identifiers. Up until now, Professor Chartreuse has gathered public data by hand by searching for keywords. Our team suggests using Perl as the programming language. Professor Chartreuse will be able to extract from the dataset using a graphical user interface. All saved data files will be deposited in cloud storage using Open Science Framework.

Open Science Framework will allow for secure access to the dataset. Professor Chartreuse will have the ability to control who has access to the data by adding and removing contributors to the project, allowing him to tailor the privacy to his team of

collaborators while sharing data with his graduate students. OSF will allow Professor Chartreuse to fork a project to share with his graduate students. Open Science Framework is open source, free to use, and has version control capabilities which will reduce the risk of losing valuable data. OSF allows contributors to tag the research data using keywords and controlled vocabularies to increase its findability and to be able to locate data quickly. This function will provide ease of data sharing between the research team and assist with reproducibility amongst other researchers using OSF. Professor Chartreuse stipulated that his data have the date, name, keywords, publications that used the data, and authors as a minimum requirement for metadata standards. In addition to having the ability to search by these queries, the team will have the ability to add a license, description of the project, and an affiliation using OSF. We recommend that Professor Chartreuse and his contributors input information in all of the metadata fields on OSF to ensure all of the metadata attributed to their projects is standardized, and therefore easy to query in OSF. In addition, it is crucial that the team uses a standard naming convention for saving all of the data files to improve searching. These steps will all contribute to ensuring quality data which will comply with the guidelines of PubMed.

We suggest that our team works with Professor Chartreuse to write ontologies that will create relationships between the concepts he is studying within the domain of SciSci. We also recommend that he install OWL2Perl on his devices. We can work with Professor Chartreuse and his collaborators to train them to have an understanding of the ontologies created and Perl's ability to parse the information from the schema, which assists in effectively extracting value from the data. This step will help to find

patterns within the dataset. Lastly, should the team require a citation software, Zotero and Mendeley are both compatible with OSF. Our data management team believes that this feature will be beneficial with sharing research within the academic community in PubMed.

While OSF's proprietary storage, called OSF Storage, allows for 5GB per individual file to be uploaded, Professor Chartreuse can choose to upgrade to an add-on cloud storage provider if his files surpass that size limit. In this case, we recommend that the team uses G Suite for Higher Education as it is free unlimited storage and will help to implement open collaboration between JCU and other institutions and is compatible with OSF. Both OSF and G Suite for Higher Education encrypt the data which will provide further security. Professor Chartreuse stated that he would like to maintain his data until he no longer works at JCU, but he may also receive up to 15GB of storage for free through Google Drive when he does decide to leave or retire.

If there is data that the research team does not need quick access to, we recommend registering the project in OSF. Once the project is registered, it is archived to OSF and will be protected as it will no longer be able to be edited or deleted. The data will be preserved by OSF; they use Rackspace as a primary server and Amazon Glacier as a backup. OSF also has a fund for preservation which will allow data to have read-access for approximately 50 years should OSF close. The research team will be producing data and disseminating it at a later date, so long-term storage is essential for reproducibility.

In conclusion, we feel that maintaining the steps of the data lifecycle can easily be accomplished with Professor Chartreuse and his research team. We want to place

emphasis in providing access to the data, which is why we suggest using Open Science Framework, as the research will contribute to the work of other researchers and will ultimately benefit science of science as a whole by enabling the data to be reused. We feel that the cloud storage option lends itself to being able to easily keep track of various contributions to the projects, while maintaining security and privacy measures. Lastly, preserving the data is a fundamental step in the data lifecycle and we feel that Open Science Framework provides enough backup servers and access control for this project.

References

BigQuery (2018) Retrieved from https://cloud.google.com/bigquery

Darwin Core (2018) Retrieved from http://rs.tdwg.org/dwc/

DigitalOcean (2018) Retrieved from https://www.digitalocean.com/

Dropbox Business (2018). Retrieved from https://www.dropbox.com/business

G Suite Business Solutions. (2018). Retrieved from gsuite.google.com/solutions/

Google cloud for higher education (2018). Retrieved from https://edu.google.com/intl/en_ca/higher-ed-solutions/

Open Science Framework (2018). Retrieved from https://osf.io/