# Data Management Plan for Professor Periwinkle

Dear Professor Periwinkle,

Below you will find a detailed Data Management Report identifying the resources required for your upcoming project on tracking marine wildlife. We hope that this plan serves you and your team well and meets all your data needs.

Sincerely,

Xx

## Purpose

The purpose of this data management plan is to make recommendations to Dr. Periwinkle on how best to achieve her long-term goal of making data from her lab available for use outside the university. Specifically, the data collected is used to write papers and to create exercises for graduate classes. However, there is currently no structured sharing of data (other than visually) from her lab outside of the university. In the future, Dr. Periwinkles goal is to make this data more accessible and available for use.

## Data Collection

### Data Type

The data collected will be that of marine wildlife to help write papers and create exercises for graduate level classes. Professor Periwinkle’s team will be using a variety of sensors and monitoring equipment to collect the data, including: i) digitally by remotely-operated marine vehicles (ROMV), ii) through tags that are surgically implanted in captured and released animals, iii) by static sensor buoys that measure ocean conditions, and iv) through communication lines that passively listen for signals for animal tags. Additional forms of data that are currently and will continue to be collected include field notes that document the animals captured and tagged, and field notes from observational studies conducted during classes Dr. Periwinkle teaches. Additionally, Dr. Periwinkle’s website allows citizen scientists to report sightings of wildlife in the Minas Basin, although this data is not currently used.

### File Formats & Size

In terms of the size and format, each collector produces data in a format specific to the sensor, however, Dr. Periwinkle’s team has developed software that converts the data to the NetCDF format for the use and storage. NetCDF is a set of software libraries and self-describing, machine-independent data formats that support the creation, access, and sharing of array-oriented scientific data[[1]](#footnote-1). The sensors are currently producing roughly 300 MB per day in raw sensor data, which becomes 500 MB in uncompressed NETCDF formatted data. It is important to note that both versions are kept. In total, the NetCDF files to date are roughly 500GB in size. Additionally, Dr. Periwinkle will be looking to use the data from the citizen scientists in the future, which is downloaded as .tsv files and is roughly 3GB. Additionally, the field notes follow the same format as the citizen scientist reports, and have a size of roughly 2GB. In sum, the size of all of Dr. Periwinkle’s tsv files amount to 500GB. It is important to note that members of Dr. Periwinkle’s team run complex simulation models that attempt to predict animal populations and movement based on oceanographic data acquired from the buoys and from collaborators. These models produce roughly 200GB of data in zipped .csv files.

### Data Categories

For the organization of the data and to help others better understand how the data is organized, Dr. Periwinkle wishes to be able to search and categorize data based on the following dimensions: organism, geographical feature, depth, and environmental conditions.

### Recommendation

It is recommended that all data collected be converted into the NetCDF file format. Although the .csv files and .tsv files are compatible to one another, it is recommended that both file types and associated data be converted to NetCDF formatted using the labs developed software. This will help ensure consistency and compatibility across all data collection mechanisms. Additionally, the use of NETCDF format will help ensure greater collaboration with other organizations in Canada as it is the primary format currently being used.

## Data Licensing

### Current Licensing

As stated, the data collected will be used to write papers and to create exercises for graduate classes. The data will be made available in visual form to the general public using OceanView.org, a platform for sharing oceans-related data. However, there is currently no structure sharing of data (other than visually) outside of the university. The intent and long-term goal is data sharing. Data sharing is the practice of making data used for scholarly research available to other investigators.

### Recommendation

It is recommended that Dr. Periwinkle use Creative Commons Attribution (CC BY) licensing in order to allow others to freely access her and her labs papers and other data without having to request access to the data files. Creative Commons is a license that is applied to a work that is protected by copyright. In essence, it is a way of easily sharing copyrighted work. CC BY licensing in particular lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered[[2]](#footnote-2). In sum, this type of licensing is recommended for maximum dissemination and use of licensed materials.

## Necessary Facilities/Equipment

### Current Facilities/Equipment

Currently, data in Dr. Periwinkles lab is primarily shared with the team using external hard drives and USB keys. Outside of the lab, as mentioned, there is currently no structured sharing of data outside of the university; researchers can request through email access to the data files, and Dr. Periwinkle can upload them to Dropbox. Additionally, the universities file sharing provider (Microsoft OneDrive) requires external users to login with a Microsoft-provided account which not all people have.

### Recommendation

It is recommended that Dr. Periwinkle, in collaboration with her institution, create an institutional repository to house all data. A data repository refers to an enterprise data storage entity into which data has been specifically partitioned for an analytical or reporting purpose[[3]](#footnote-3). Specifically, an institutional repository is a type of digital library that captures the original research and other intellectual property generated by an institutions constituents. Institutional repositories are generally maintained by academic libraries at higher education institutions and offer external individuals the ability to participate and access open data[[4]](#footnote-4). A benefit of create an institutional repository is that Dr. Periwinkle will be able to create separate repositories for each of her required categories. This will better help organize the data based on defined characteristics. Additionally, her team will all be able to access and share their relevant data through one common entry point, instead of relying solely on USB keys and external hard drives.

The data housed in the institutional repository will be accessible to the public without having to formally request access. Ultimately, this will allow greater data sharing. Shared data enables collaboration, builds a shared understanding, helps anticipate future problems, and leads to better research practices. Additionally, the repository will be able to accommodate the recommended licensing, which is CC BY. This will enable Dr. Periwinkle to create her own platform for sharing oceans-related data, which can be built to suit the required size of all the data files. The institutions librarian has agreed to help Dr. Periwinkle and her lab build the repository. For further examples, it is suggested that Dr. Periwinkle Utilize Canadian Research Data Management (RDM) resources, such as the *Canadian Association of Research Libraries[[5]](#footnote-5)* and *Protage Network[[6]](#footnote-6)* for best practices and to see other Canadian institution repository examples.

## Recommended Data Management Practices

### Current Data Management Practices

Data has been collected in some capacity in Dr. Periwinkle’s lab since 1998; however, over the last three years she is now collecting substantially more data than ever before (for example, the ROMV are collected on a monthly basis).Currently, past data and current data (as described above) is kept in Dr. Periwinkles office in a myriad of forms including: floppy disks, ZIP Disks, CDs, DVDs, Blu-Ray, external hard drives, USB keys as well as through Dropbox.

### Recommendation

It is recommended that Dr. Periwinkle move to a more consistent means of backing up current data and archiving past data. The establishment of the institutional repository will serve as a strong platform for Dr. Periwinkle to archive past data. With this, it is recommended that Dr. Periwinkle convert and transfer her data files from all the forms mentioned above to the repository. Major funding agencies worldwide are pushing researchers to archive their data and make it available to other scientists. Re-collecting data is inefficient, therefore making it available to the public in an accessible manner is highly recommended.

Although data repositories can also serve as backups during rare events where data are lost, it is still recommended that Dr. Periwinkle perform her own data backup on an external hard drive. Hard drive costs per gigabyte have dropped substantially over the years. Because Dr. Periwinkle currently roughly 1 terabyte of data, and is scheduled to keep collecting more data, it is recommended she purchase an external hard drive that can house roughly 5 terabytes. This will cost Dr. Periwinkle less than $200.00 at a one-time expense[[7]](#footnote-7) and will be than enough storage space. As Dr. Periwinkle’s data continues to grow, she can look to purchase additional external hard drives to help back up her data.

## Ownership and Access to Data

### Current Ownership

Currently, there is no procedure or contract in place that the students (incoming or leaving) have to sign when working in Dr. Periwinkles lab. As students graduate, they can share their data with current students in the lab or take it with them on their personal devices. In terms of ownership, Dr. Periwinkle owns the data she and her lab produces, however, the data produced by citizen scientists are open.

### Current Access

In addition to the lab members having access to the data, the data is made available in visual form to the general public using OceanViewer.org. However, as mentioned, there is currently no structured sharing of data outside of the university. Therefore, those outside of the university have to request access to Dr. Periwinkles data. Although the university uses the file sharing provider Microsoft OneDrive, it requires external users to have an account, which not everyone does.

### Current Data Sharing Agreement

There are other research teams in Canada and around the world that collect data that might be interesting to include in Dr. Periwinkle’s team’s research; the lab currently requests access to this data through personal connections and relationships. Organization’s as well as individual research groups in Canada collect similar data. However, there is currently no data sharing agreement in place, merely a collection of ad hoc sharing opportunities.

### Recommendation

In terms of ownership, it is recommended that Dr. Periwinkle create a contract for all students to sign around ownership, confidentially, and legality to ensure consistency and credibility amongst her team. Additionally, through the CC BY licensing, it is recommended that Dr. Periwinkle take ownership of her labs data.

With the recommended creation on an institutional repository, this will allow the general public to have access to all documents, alleviating the need for data sharing. Additionally, the repository will allow external users to request to deposit items related to the relevant topics while maintaining their own ownership of their data.

It is recommended that a data sharing agreement be made between the other research teams and organizations in Canada in order to effective collaborate and share data. In this data sharing agreement, it is recommended that all constituents agree to submit all data to the Ocean Biogeographic Information System (OBIS) as it is the open-access data and information clearing-house on marine biodiversity for science, conversation and sustainable development in Canada and across the Globe. Additionally, through the creation of the repository and in congruence with the data sharing agreement, Dr. Periwinkles lab will have greater access relevant research in Canada and beyond. It is also important to note that Dr. Periwinkle has indicated that any research that is contributed outside of the English language will be translated, therefore, it serves as a great platform to garner research and information from all over the world.

## Post-Project Data Value, Reuse, and Long-Term Preservation

### Value

As mentioned by Dr. Periwinkle, all data is valuable. In order to help increase the long-term value of the data, it is recommended that the lab’s 2003 help document be revamped to include the proposed Data Management Plan recommendations. Currently, the help document describes how to work with the data sets, and how to contribute to the datasets. In addition to these sections needing to be updated, the document should also include how to convert the necessary files to NetCDF, to ensure consistency. Additionally, the document should be shared and housed in a universal location, that is accessible by all lab members.

With the implementation of the institutional repository, all data will be able to retain value after the life cycle of the project. Specifically, storing data in data repositories is highly encouraged and is part of the preservation portion of the data lifecycle. Data repositories can help make a researcher’s data more discoverable and accessible, and lead to potential reuse.

### Reuse

In order to maximize the re-use of the data, fellow researchers should utilize the intuitional repository in order to perform follow-up research. Additionally, any new research should also be added to the repository. Dr. Periwinkle should also undertake research reviews, scrutinizing the findings and helping her students learn and build off of findings. In addition to the repository, the data will be submitted and made available to others through the OBIS. This in turn will ensure reusability.

### Preservation

In order to preserve the data, Dr. Periwinkle must migrate the data to the NetCDF format. Following this, the lab must migrate the data to the suitable medium, which is the recommended institutional repository under the appropriate repository category. The data should consistently be backed-up on the external hard drive to ensure its safety. Following this, Dr. Periwinkle will need to create metadata and documentation, and then archive the data (which can be done through the repository and through an external hard drive).

## Metadata Requirements & Relevant Linked Open Data Strategies

### Metadata

Metadata is data that provides information about other data[[8]](#footnote-8). The field notes follow the Darwin Core. Darwin Core is a body of standards for biodiversity informatics. In essence, it provides the stable terms and vocabularies for sharing biodiversity data. Darwin Core It is important to note that OBIS has been transitioned globally to the Darwin Core. All the required files must be filled, otherwise the data are dismissed. The Darwin Core guidelines that encompass a series of terms related to the following classes: Taxon, Identification, Occurrence, Record Level, Location, Event, and Material Sample[[9]](#footnote-9). This should be followed by Dr. Periwinkle and her lab.

### Linked Open Data Strategies

Dr. Periwinkle should look to link any and all relevant data in the data sets.

## Storage Costs

### External Hard Drive

As noted, the cost of an external hard drive as a backup mechanism will cost Dr. Periwinkle around $200.00 per drive. This will be essential to ensure the safety of the data.

### Institutional Repository

It is recommended that Dr. Periwinkle and her institution use the platform *DSpace* to build the repository. DSpace is a digital repository system with the ability to capture, store, index, preserve an redistribute an organization’s research material in digital formats. Research institutions worldwide use DSpace; DSpace is freely available as open source[[10]](#footnote-10). DSpace is written in Java, and will run on any operating system. For the convenience of the user, DSpace packages the necessary JDBC and other drivers and libraries together.

In particular for Dr. Periwinkle, it is recommended she purchase

**Professor Pinkerton**

Necessary Data

Licensing Information

Necessary Facilities/Equipment

Recommended Data Management Practices

Ownership and Access to Data

Post-Project Data Value

Metadata Requirements

Relevant Linked Open Data Strategies

Reuse and Long-Term Preservation

Storage Costs

**Professor Chartreuse**

Necessary Data

Licensing Information

Necessary Facilities/Equipment

Recommended Data Management Practices

Ownership and Access to Data

Post-Project Data Value

Metadata Requirements

Relevant Linked Open Data Strategies

Reuse and Long-Term Preservation

Storage Costs

1. unidata. (n.d.). Network Common Data Form. In *unidata*. Retrieved from https://www.unidata.ucar.edu/software/netcdf/ [↑](#footnote-ref-1)
2. Creative Commons. (n.d.). About the Licenses. In *Creative Commons*. Retrieved from https://creativecommons.org/licenses/ [↑](#footnote-ref-2)
3. Informatica. (n.d.). What is a data repository?. In *Informatica*. Retrieved from https://www.informatica.com/ca/services-and-training/glossary-of-terms/data-repository-definition.html#fbid=m62zCwY0K\_4 [↑](#footnote-ref-3)
4. Burns, C., Lana, A., & Budd, J. (2013). Institutional Repositories: Exploration of Costs and Value. *D-Lib Magazine*, *19*(1). doi:10.1045 [↑](#footnote-ref-4)
5. http://www.carl-abrc.ca/ [↑](#footnote-ref-5)
6. https://portagenetwork.ca/ [↑](#footnote-ref-6)
7. https://www.bestbuy.ca/en-ca/category/external-hard-drives/20237.aspx [↑](#footnote-ref-7)
8. Merriam-Webster. (2018). Metadata. In *Merriam-Webster*. Retrieved April 8, 2018. [↑](#footnote-ref-8)
9. Ocean Biogeographic Information System. (n.d.). Darwin Core Manual. In Ocean Biogeographic Information System. Retrieved from http://www.iobis.org/manual/darwincore/ [↑](#footnote-ref-9)
10. DSpace. (2015). About DSpace. In *DuraSpace*. Retrieved from https://wiki.duraspace.org/display/DSPACE/User [↑](#footnote-ref-10)