Group Project !

**Professor Periwinkle**

# Data Management Plan for Professor Periwinkle

Dear Professor Periwinkle,

The purpose of this data management plan is to make recommendations to you on how best to achieve your long-term goal of making data available for use outside the university. Additionally, we are aware that 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 your lab outside of the university.

With your goal to make this data more accessible and available for use, we recommend submitting all pertinent data files to the Ocean Biogeographic Information System (OBIS) repository. This will help make your data accessible and usable by all in a consistent and convenient manner. OBIS accepts data from any organization, project, or individual who wants to contribute data. Moreover, you will remain the owner of your data. Additionally, the OBIS follow’s the Darwin Core, which are the data standards currently implemented at your lab.

In terms of internal data sharing, we recommend that you utilize your institutions current data sharing platform, Microsoft OneDrive. This will ensure that all data is properly stored and in one location, rather than on external hard drives and USB keys.

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

Sincerely,

Xx

## 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’s lab move to using the universities file sharing provider, OneDrive, when working with their data in addition to using external hard drives and the USB keys. This will ensure that all data and files are housed in one location and will act as a means for backup. In terms of making data accessible and available to external users, it is recommended that Dr. Periwinkle upload her files to the Ocean Biogeographic Information System (OBIS) repository. 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). These repositories are digital libraries that capture the original research and other intellectual property generated by an institution’s constituents. This repository would be maintained by OBIS and would offer external individuals the ability to participate and access open data.

The data housed in the OBIS 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, as all data is owned by individual or group themselves. With this, it is recommended that Dr. Periwinkle convert and transfer the necessary data files from 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.

## 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. As mentioned, moving to the OneDrive platform will serve as a means to backup the data. In addition to OneDrive, the OBIS repository will serves will serve as a strong platform for Dr. Periwinkle to archive current and past data from the above mentioned platforms.

Although data repositories can 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[[4]](#footnote-4) 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 recommendation to submit data to OBIS, this will allow the general public to have access to all documents, alleviating the need for data sharing.

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 (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 this 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, especially since OBIS is an international platform.

## 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 on OneDrive where it can be accessible by all lab members.

With the submission of Dr. Periwinkles data files and documents to OBIS, 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 OBIS in order to perform follow-up research. Additionally, any new research conducted by Dr. Periwinkle and her lab should also be added to the OBIS repository. Dr. Periwinkle should also undertake research reviews, scrutinizing the findings and helping her students learn and build off of findings. 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 OBIS repository under the appropriate repository categories. The data should consistently be backed-up OneDrive and 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[[5]](#footnote-5). 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[[6]](#footnote-6). 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. As mentioned, 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. These can be used to link her data from the OneDrive to the OBIS repository and allow for easier querying.

## 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.

### Microsoft OneDrive

As mentioned, Microsoft OneDrive is the current platform used by the University. With this, Dr. Periwinkle and her lab will be able to store and access their data on this platform for free under the universities agreement.

### OBIS Repository

As mentioned, it is recommended that Dr. Periwinkle and her institution use the OBIS to submit and share data. This platform is free to use and will allow Dr. Periwinkle to keep ownership of her data. It is important to note that OBIS only harvests data from recognized OBIS nodes. If Dr. Periwinkle wishes to become a node, she can do so through the application below. OBIS nodes are responsible for representing all aspects of OBIS within a particular region or taxonomic domain. For further information and to register with OBIS, please see the links below.

**OBIS Canada**

Centre of Marine Biodiversity, Bedford Institute of Oceanography

<http://www.marinebiodiversity.ca/>

**Tobias Spears**

[tobias.spears@dfo-mpo.gc.ca](mailto:tobias.spears@dfo-mpo.gc.ca)

**Mary Kennedy**

[kennedym@mar.dfo.mpo.gc.ca](mailto:kennedym@mar.dfo.mpo.gc.ca)

**Node Application**

[Application Form](https://iode.org/index.php?option=com_oe&task=viewDocumentRecord&docID=11793)

**Professor Pinkerton**

**Professor Pinkerton**

Introduction and Recommendation

Professor Pinkerton has a very large, and ever growing, collection of data that is used by herself and her coworkers. Some of her files contain open data and she would like everyone to have access to them, while some contain sensitive information with access awarded on a case-by-case basis. This data management plan attempts to provide a way for Professor Pinkerton to easily organize her data sets so that they can be used in the future, either by herself or other individuals.

Given the vast number of spreadsheets that Professor Pinkerton currently has (and the amount that she is likely to collect in the future), it is our recommendation that a database management system (DBMS) is set up to organize this data. Contained within this DBMS will be the metadata about all of the different spreadsheets that she has collected. This DBMS will help Professor Pinkerton keep a record of all of the data sets that she has in an organized fashion, as well as help her locate them easily when needed. It will primarily operate as a database with information about all of the data sets that Professor Pinkerton collects. This DBMS should be available online so that anyone with data requests can easily find what data Professor Pinkerton has. The DBMS should also include links to where the data sets can be downloaded so that they can easily be accessed. For the data sets that are not open, the link will bring them to a page where they can request access to the set from Professor Pinkerton via email, at which point if she wants to grant them access she can do so.

We are recommending that all of the spreadsheets should be hosted on the university drive (DalSpace), with a hard disk back up. The files will have to be separated in to two different categories (with the category of data being noted in the metadata, and therefore in the DBMS). The first category of spreadsheet is for open data, which anyone can have access to. The second category of data is for files that Professor Pinkerton would like to control who has access to.

The spreadsheets that are in the first category will be on a drive location that is open and accessible to anyone, while the spreadsheets that are in the second category will be stored on a separate location on the drive, and each of these files will be protected so that only people that Professor Pinkerton has given permission to can access them (by responding to the requests and giving them access). People will have to request permission to access these spreadsheets. Both of these drives will be protected so that no one, other than Professor Pinkerton and potentially her postdoc students (at her discretion) can upload or modify files that are on the drive. This will prevent any changes from being made to the raw data sets that are a part of her collection.

Necessary Data

There is no new data that Professor Pinkerton actually needs to collect as a part of this data management plan. Rather, as her collection of data increases and she gets more and more spreadsheets, these new sheets will be integrated into her collection based on the database management system that gets set up. The metadata will need to be entered into the DBMS as the spreadsheets are added to the collection, and will need to be uploaded to the appropriate drive location (depending on whether the data is open or not), and backed up on the local backup.

Professor Pinkerton does need to locate and organize the metadata about the different data sets. This is all information that she currently has, it will just need to be compiled into the DBMS. The only new piece of information is the links to where the individual data sets are stored on the DalSpace drive.

Licensing Information

The licensing information for each of the spreadsheets needs to be contained in the metadata, and therefore noted in the DBMS. Each spreadsheet will have its own licensing requirements, and will therefore need to be tagged as either copyright, copyleft, or creative commons (BY, NC, ND or SA). Whichever licensing is on the data sets when Professor Pinkerton obtains them, needs to be included in the DBMS.

Necessary Facilities/Equipment and Costs for Data Storage

It is our recommendation that all of the data be stored in two locations. The first location is on the university cloud that is used at Dalhousie University (DalSpace). The second location is to have local backups of the data on the hardware where the files are currently stored. As a result, no new facilities or equipment are required to organize and maintain the data sets.

Because of the vast amount of data that Professor Pinkerton currently has, and the rate at which she is continuing to compile data, more storage may be necessary in the future. This would require both an expansion of the cloud that Dalhousie uses, as well as more hard drive space to store the data locally. Because Dalhousie University already has both an online storage drive, and hardware for backing up this information, no additional costs will be incurred to store this data.

We also recommend that the DBMS is created using a program that makes use of a search query language (SQL). One program that she should consider using is SQLite, which can be downloaded for free, and thus again resulting in no additional costs. This DBMS should be backed up both on the cloud and locally, similar to the other data sets.

Ownership and Access to Data

The DBMS would be owned by Professor Pinkerton, however should be available to anyone that is looking for data sets. Professor Pinkerton can provide a link to the DBMS to anyone looking for data so they can locate what they are looking for, without having to request the data sets from her by email (except for the protected sheets).

As mentioned before, the spreadsheets that Professor Pinkerton has will need to be separated into two different categories. The different categories will have different access and ownership characteristics. For data that is considered to be open, the spreadsheets should be made publicly available on the Dalhousie University online drive. This will reduce the amount of requests that Professor Pinkerton will receive to send out her data sets as people can download these data sets themselves.

For the second category of data sets that are not considered to be open, the access will be issued on a case-by-case basis. These will be stored on a location of the drive that is not shared. For someone to access one of the data sets, they will need to seek permission from Professor Pinkerton, and she will send them the link to access the specific data. This will also allow her to turn off link sharing of any of the files when she no longer wishes to share the file with others.

Post-Project Data Value

By setting up this DBMS, significant value is obviously created. Once the data sets are hosted online and interested parties have access to the DBMS, the data sets will be much easier to find for interested parties. Additionally, it will be easy to add new metadata for the DBMS as new data sets are collected. It will also reduce the number of data requests that Professor Pinkerton will receive, as interested parties will be able to download the data that they need (as long as it is tagged as open data). This DBMS also helps Professor Pinkerton comply with the FAIR data principles, adding further value. The DBMS will help with all aspects of the FAIR principles (findable, accessible, interoperable and reusable).

Metadata Requirements

The inclusion of accurate and thorough metadata is critical for maintaining the DBMS in an organized fashion. The metadata that needs to be included about each of the spreadsheets should be:

* Title
* Subject
* Size
* Original Author
* Licensing Information
* Accessibility
  + Whether the file is Open or has restricted access
  + Who has access to the file
    - For files tagged as open, the default should be that anyone has access to the file
* Date
* Size
* Link to Drive Location

This data should be entered into the recommended SQLite program for each new data set that is collected. The metadata should be accurate and complete for each set to ensure that a correct record of all of the data sets that Professor Pinkerton owns are logged and can be accessed by others with ease. Below is the entity-relationship model, which demonstrates the pieces that make up the metadata, and how they are related. Important features of the entity-relationship model are that each data set can have multiple authors and subjects so that the data sets can be easily queried.

Overall, by creating a DBMS, Professor Pinkerton can ensure that her data is organized and easy to use for both herself, and others in the future. By storing the data both on the Dalhousie University drive, and having local backups, Professor Pinkerton can be confident that she will have this data for the foreseeable future.

**Professor Chartreuse (Priyanka) – This is the datalord**

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. https://www.bestbuy.ca/en-ca/category/external-hard-drives/20237.aspx [↑](#footnote-ref-4)
5. Merriam-Webster. (2018). Metadata. In *Merriam-Webster*. Retrieved April 8, 2018. [↑](#footnote-ref-5)
6. Ocean Biogeographic Information System. (n.d.). Darwin Core Manual. In Ocean Biogeographic Information System. Retrieved from http://www.iobis.org/manual/darwincore/ [↑](#footnote-ref-6)