# Data Management Plan for Professor Periwinkle

Dear Professor Periwinkle,

Below you will find a detailed Data Management Report (DMR) 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 collected include field notes that document the animals captured and tagged, and field notes from observational studies conducted during classes Dr. Periwinkle teaches. Currently, Dr. Periwinkle’s website also 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 (UNIDATA). The sensors are currently producing roughly 300 MB per day in raw sensor data, which becomes 500 MB in uncompressed NETCDF formatted data. In total, the NetCDF files to date are roughly 500GB in size. It is important to note that both versions are kept. 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 500GB. Additionally, the field notes follow the same format as the citizen scientist reports, and have a size of roughly 2GB. 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 using the following dimensions: by organism, geographical feature, depth, and environmental conditions.

### Recommendation

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 in use.

## 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 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. In sum, this type of licensing is recommended for maximum dissemination and use of licensed materials (CREATIVECOMMONS).

<https://creativecommons.org/licenses/>

## 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. Additionally, as mentioned, there is currently no structured sharing of data outside of the university and 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 (INFORMATICA). 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 area.

The data housed in the institutional repository will be accessible to the public without having to formally request access. 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 should have no problem helping Dr. Periwinkle set up the repository. For further examples, it is suggested that Dr. Periwinkle visit *Canadian Association of Research Libraries* and *Protage* for best practices and to see other Canadian institution repository examples.

<https://www.informatica.com/ca/services-and-training/glossary-of-terms/data-repository-definition.html#fbid=m62zCwY0K_4>

## 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 monthly.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, 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. 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 of current data. Hard drive costs per gigabyte have dropped substantially over the years. Because Dr. Periwinkle currently has over 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.

## Ownership and Access to Data

**• Who will own and have access to the data**

### Current Ownership

Currently, student data is open and there is no procedure or contract in place that the students (incoming or leaving) have to sign. As students graduate, they can share their data with current students in the lab, or take it with them on their personal devices. Additionally, as students graduate they share their data with current students in the lab, or take it with them on their personal devices. In contrast, Dr. Periwinkle owns the data she produces.

### Current Access

In addition to the lab having access to the data, the data is made available in visual form to the general public using OceanViewer.org. However, 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.

### Recommendation

In terms of ownership, it is recommended that Dr. Periwinkle create a contract for all students around ownership, confidentially, and legality, to ensure consistency and credibility amonst her team.

## Post-Project Data Value

**• Which data will retain value after the life of the project**

## Metadata Requirements & Relevant Linked Open Data Strategies

**What metadata and linked open data strategies will be employed**

The field notes follow the Darwin Core: <http://www.iobis.org/manual/darwincore/>  
All the required filed must be filled, otherwise the data are dismissed.

Descriptive: describes a resource for purposes such as

discovery and identification.

• Structural: is about containers of data and indicates how

compound objects are put together.

• Administrative: provides information to help manage a

resource.

## Reuse and Long-Term Preservation

**How will its reuse be enabled and long-term preservation ensured after**

**the original research is completed**

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.

follow up research

new research

undertake research reviews

scrutinize findings

teach and learn

## Storage Costs

**• How much will the storage of this data cost (cloud and/or hard drives)**

Dr. Periwinkle has funding in her grant for the Data Management Plan. From Innovation Canada.

<https://www.unidata.ucar.edu/software/netcdf/>

**Our Recommendation**

As stated, the data collected is used to write papers and to create exercises for graduate classes, however, there is currently no structured sharing of data outside of the university. The long-term goal is to make this data available for use. In order to do so, in conjunction with the information from this data management plan, we recommend that Dr. Periwinkle …

Utilize Canadian Research Data Management (RDM) resources, such as Portage Network and Research Data Canada.

Shared data enables collaboration, builds a shared understanding, helps anticipate future problems, and leads to better research practices.

Research Data Management allows for digital data to be stored and management electronically. It is easy to share, replicate, and combine with other data. However, the data must be contained and maintained so they are consistent with the goal of long-term preservation.

**Current Issues**

Someone looking to access Dr. Periwinkles work can email Dr. Periwinkle to request the data files, and Dr. Periwinkle can upload them to Dropbox. It is important to note that although the university uses the file sharing provider Microsoft OneDrive (a cloud service), it requires external users to login with a Microsoft-provided account of some kind and not all people have these accounts.

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; they currently request access to this data through personal connections and relationships. Organization’s in Canada collect similar data and there are also individual research groups collecting data on specific species, specific sites, etc. However, there is not currently a data sharing agreement in place, merely a collection of ad hoc sharing opportunities. In contrast, NOAA in the US has the IOOS system which spans the country, and the European Commission has EMODnet.

Many funding agencies and institutions have policies on data sharing, because openness and transparency are fundamental to the scientific method.

Everyone who starts with Dr. Periwinkle’s team is told how the data is formatted and structured. There is a help document from 2003 describing how to work with the data sets, and how to contribute to the datasets; this help document is on a shared drive somewhere. As students graduate they share their data with current students in the lab, or take it with them on their personal devices.

DATA COLLECTED OUTSIDE OF CANADA MAY BE IN A DIFFERENT LANGUAGE. DR PERIWINKLE IS PLANNING ON HIRING A TRANSLATOR FOR THOSE CASES

**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