Data Management Plan

Presented to Dr. Periwinkle

**Executive Summary**:

Outlined in this report is a data management plan, which includes recommendations for database management, data storage and data sharing. It is understood that there has been a considerable amount of data collected and shared amongst your team and other interested parties since 1998.

Best practices and policies of government funding agencies call for data to be made available as *open data*, and follow the FAIR (findable, accessible, interoperable and reusable) guiding principles for data management. This is not only part of the requirement for government sourced funding but it also is a growing part of data management best practices. It meets the university’s policy on research data, is often required by peer reviewed journals, and is shows accountability for public funds.

We have prepared this report - based on requirements outlined by the university librarian and input from the research project manager, Dr. Periwinkle - to be thorough and of benefit to the research team.

# **Project Description**

* Project is funded by innovation canada, and research excellence fund.

Professor Periwinkle and her team of researchers have been monitoring marine wildlife using a variety of sensor techniques, including tags and arrays. This research collects copious amounts of raw sensor data which is converted to Network Common Data Form (NetCDF) format. This research is also supplemented with the generation of complex simulations run from the data, saved as csv files. Looking forward, Dr. Periwinkle and her team are want to be able to share this research with other research teams as well as provide a means of sharing contributions between members. The envisioned collaborative space would be similar to the Integrated Ocean Observing System (IOOS) managed by NOAA in the United States, or the European Marine Observation and Data Network (EMODnet). Additionally, it is the hope of researchers to share data through the Ocean Biogeographic Information System (OBIS), one of the world's leading databases for ocean biodiversity and biogeographic data (OBIS, N.D.).

# **Documentation, Organization, and Storage**

Based on consultations with the research team, it is understood that along with raw sensor and array data, researchers record field notes using *Darwin Core*, and then convert these file to *NetCDF* format, as well as run simulations are saved in .csv file type. That being said, other contributors are not required to use this format. It is also estimated from the information provided that approximately 705 gigabytes (GBs) of data has been collected in the past and this will continue to grow at approximately 800 megabytes (MBs) per day or 24 GBs per month. This value is displayed monthly as it has been explained that data is collected on a monthly basis.

In order to properly store and backup this valuable data, it is proposed that a local storage server be created in addition to a cloud based storage solution. The local storage server should utilize a RAID, or Redundant Array of Inexpensive Disks, in the form of physical hard drives. This system adds a layer of data redundancy by storing the files across multiple drives which protects against failure (Jacobi, 2012). For an added layer of security, it is recommended that the drives be organized using RAID 5, and that three 5 Terabyte (TB) hard drives are used. RAID 5 requires that three hard drives be used and this ensures that data is backed up across all of the drives. In this configuration, RAID 5 will allow for 5 TBs of storage as it utilizes data parity across all of the drives (Jacobi, 2012). With the price of storage continually declining, this storage solution is expected to cost $150 per hard drive and $100 for the hard drive enclosure. This would bring it to a grand total of an estimated $550.

In terms of a cloud software solution, this will be required so that citizen scientists and global research teams are able to collaborate and share data easily. Since the research and infrastructure is funded by Innovation Canada and the Research Excellence Fund, this limits some storage options. In accordance to the Government of Canada’s Cloud Adoption Strategy, this requires that data under government control be stored on servers that reside in Canada (Government of Canada, 2016). Based on this requirement, it is proposed that the cloud storage software, Sync.com, be utilized by the team. This software provides users with Canadian server storage, end to end encryption, and ample storage for users all at an affordable price (Sync.com, 2018). Per user, each can store up to 10 terabytes of data, which will allow researchers to continue to add data, at the same rate, for over years to come. Each user license costs $15 per month and members are able to set up a free account (with limited functionality) which would be ideal for other researchers looking to download the data (Sync.com, 2018). This places the cost of online storage at $540 per year for three licenses.

Once these pieces of equipment of have been acquired, it will be important that all existing data, such as those stored on floppy disks, CDs, and DVDs, be uploaded to these databases.

Going beyond the equipment required to store the data, there is a great deal of importance that also lies in both the data itself and file naming conventions. As mentioned through conversations, it is the goal of researchers to share data with the Ocean Biogeographic Information System (OBIS) at some point in the future. With that in mind, it is recommended that the data standards of OBIS be implemented at this time. This is particularly important with regards to three pieces: the species occurrence data, dataset metadata, and sampling specific data (OBIS, N.D.).

With regard to species occurrence data, the use of Darwin Core should continue to be used. Dataset Metadata however should be configured to follow Ecological Metadata Language (EML). This metadata standard was developed in cooperation between OBIS and the Ecological Society of America and associated efforts, and has a minimum requirement of four terms: Title, citation, contact, and abstract (OBIS, N.D.). Finally, for sampling specific data, like organism, geographic feature, depth, or other interesting metrics, Darwin Core Archive and OBIS-ENV-DATA standards should be used (OBIS, N.D.). This archive system acts as a star schema with each of the sample data pieces like organism, as ID columns. Using this Archival standard will make searching the data very easy and highly standardized as you will be able to search for specific ID columns.

When applying this information to the FAIR model of research data management having these storage solutions and data standards insure that the data is both Findable and Interoperable.

# Access, Sharing and Re-use

The ultimate goal of utilizing the recommended data storage and data standards is to both ensure that a backup of the data is created as well as provide a means of collaboration and sharing within teams and external parties. Although citizen scientists’ data follow open data policies, appropriate licenses should be applied to publicly funded research. This is done to protect the intellectual property of the researchers and adds accountability to the research.

The Canadian Government provides funding to researchers through several funding agencies. It is understood that the research and infrastructure of Dr. Periwinkles team is funded by Innovation Canada and the Research Excellence Fund. Both of these funds are provided through various bodies within the Canadian Government and thus have policies in place for the Data.

In accordance to the rules of The National Research Council (NRC), researchers who are federally funded are required to have peer-reviewed journal publications made freely available after 12 months of publication (NRC Website.). In order to meet this requirement, researchers can either archive their full text manuscript online, or publish in a journal that offers immediate access ([www.cdnsciencepub.com/our-journals/open-access/quick-guide-to-oa.aspx#3](http://www.cdnsciencepub.com/our-journals/open-access/quick-guide-to-oa.aspx#3)).

To best protect the intellectual property of research teams, it is recommended that a creative commons licences be applied to the work of Dr. Periwinkle and team. Specifically, the license should be a creative commons attribution required Sharealike (CC-BY-SA). This license requires that users of the data must give appropriate credit, provide a link to the license, and indicate any changes that were made to the data. Additionally, the Share Alike element of this license ensures that if another person were to transform or build upon the original material they must also distribute this material using the same license. By applying this license to the research being conducted it ensures that the intellectual property of the data is not lost while still ensuring that it is available.

Currently, research assistants and students that have access to the teams data are not required to sign a contract indicating that the data and findings are owned by Dr. Periwinkle. In addition to applying licenses to the data, it is also recommended that a contract be instituted to ensure that ownership of the data is clear and legally binding.

Assuming that local storage and cloud storage repositories will be created for the data, this ensures that both team members and interested parties will have available access to the data.

Following a generic data lifecycle, Data starts at the ‘creation’ phase and ends at a phase of ‘re-use’ after the life of the project and data is no longer being added. In order to ensure that this ‘re-use’ phase is reached, it is important that the data be properly preserved and able to be accessed. It is assumed that all of the data will retain its value after the life of the project given that there are many insights that can be obtained through sensor data and field notes. It is imperative that this research be preserved so that future research can be conducted using it and it can be used to inform new research endeavours.

# Archiving

* How will data be archived? Is it going to be stored in an archive for long-term access? If not how will it be preserved?
* Is this a good place to include a couple of statements about OLAP/OLTP?
* OLTP - Online Transaction Processing = “Live” data
* OLAP - Online Analytical Processing = Historical data being actively analyzed

As previously mentioned, it is the goal of the Dr. Periwinkle to eventually share the data through OBIS’ database archives.

# Conclusions and Recommendations

Based on the information that you provided it is understood that the sharing of data is important. That being said, it will also be important to ensure that those who contributed the work are recognized. It is recommended that a Creative Commons BY (CC BY) license be applied to the data collected and shared. This license ensures that team members and other contributors can still access, modify, and commercialize their work, however, the original creator must be recognized. This license type provides users with complete freedom of the data while also protecting intellectual property.

From Elvira’s PPT’s - week 10:

Data Management Plan Elements

• Data Collection

• Documentation and Metadata (3 types:descriptive, structural & administrative)

• Storage and Backup

• Preservation

• Sharing and Reuse

• Responsibilities and Resources

• Ethics and Legal Compliance

We see to have covered everything, except the ethics and legal compliance.

# References:

Jacobi, J.L. (2012) RAID made easy. *PCWorld.com*. Retrieved from [www.pcworld.com/article/194360/raid-made-easy.html](http://www.pcworld.com/article/194360/raid-made-easy.html)

Treasury Board Secretariat. (2016). *Government of Canada Cloud Adoption Strategy*. Retrieved from [www.canada.ca/en/treasury-board-secretariat/services/information-technology/cloud-computing/government-canada-cloud-adoption-strategy.html](http://www.canada.ca/en/treasury-board-secretariat/services/information-technology/cloud-computing/government-canada-cloud-adoption-strategy.html)

N.A. (2018) Why Sync? *Sync.com*. Retrieved from: [www.sync.com/your-privacy/](https://www.sync.com/your-privacy/)

N.A. (2018) Pricing. *Sync.com*. Retrieved from: [www.sync.com/pricing/](https://www.sync.com/pricing/)

N.A. (N.D.) How to Contribute data to OBIS? *IOBIS.org*. Retrieved from: [www.iobis.org/manual/#contribute](http://www.iobis.org/manual/#contribute)

N.A. (N.D.) Ecological Metadata Language. *IOBIS.org*. Retrieved from: [www.iobis.org/manual/eml/](http://www.iobis.org/manual/eml/)

N.A. (N.D.) Vision & Mission. *IOBIS.org.* Retrieved from: <http://www.iobis.org/about/>