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**

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).

# **Documentation, Organization, and Storage**

* type of data collected, facilities and equipment required, cost of storage,
* metadata linked and which open data strategies employed.
* estimated size of storage (see question about formatting and naming conventions).

Based on consultations with the research team, it is understood that along with sensor and array data, researchers record field notes using *Darwin Core*, and then convert these file to *NetCDF* format, as well as 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 you provided that approximately 1.2 Terabytes (TB) of data has been collected in the past and this will continue to grow at approximately 300 gigabytes (GB) per year.

Although this sounds like a lot of data, the price of physical storage of hard drives is ever declining. This means that the purchase of a 5 TB hard drive can be had for less than $200 CAD. With this amount of storage, and velocity of data collection, it would last researchers more than 10 years. In addition to this physical storage and backup, it will also be important to supplement the data storage with a cloud-based solution for easy transfer between team members and also exterior contributors. INSERT CLOUD BASED STORAGE PRODUCT HERE

# Access, Sharing and Re-use

The Canadian Government provides funding to researchers through several funding agencies. The National Research Council (NRC) spells out how to meet their policy on open data on their website:

Canadian funding agencies (CIHR, NSERC, SSHRC), have recently adopted [**a new policy**](http://www.science.gc.ca/default.asp?lang=En&n=F6765465-1) that “requires federally funded peer-reviewed journal publications to be made freely available within 12 months of publication...”

This requirement can be met by doing one of the following:

* Grant recipients **archive the final peer-reviewed full-text manuscript in an online repository** where it will be freely accessible within 12 months
* Grant recipients can **publish in a journal that offers immediate open access** or that offers open access on its website within 12 months.

NOTE: The two options above are not mutually exclusive.

([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))

* Include: what licenses apply, who will own and have access, which data will retain value after the life of the project, How will its reuse be enabled and long term preservation ensured after the original research is completed. Provide example of software they could use.

I think the Creative Commons Licenses that would apply to publicly funded research would be **CC BY-SA** (Attribution — You must give appropriate credit, provide a link to the license, and indicate if changes were made & ShareAlike — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.)

* There’s an online CC License ‘chooser’ that generates the correct license for each use - very useful (<https://creativecommons.org/choose/>)
* An additional way to share the accumulated data would be Wikidata. We can mention this as a method of use and re-use of the data for researchers around the world - although I don’t know how all that data gets ported over to wikidata

# 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

# Conclusions and Recommendations

* Just like toast masters, make sure to tell them what you’re going to tell them, tell them, and then tell them again in summary.

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.

Test Test