INFO6540
Assignment 3

Elvira Mitraka
20180410

DATA MANAGEMENT PLAN: MARINE WILDLIFE AND OCEANS DATA

Consultants: **F.A.M.E Data Management –** for data that are **F**indable, **A**ccessible, **M**emorable, & **E**nduring. Team members: Jenny Farbstein, Lisa Nardecchia, Victoria Roberts, Carlye Stein

Principle Investigator: Professor Periwinkle

Funder(s): Innovation Canada and the Canada First Research Excellence Fund

Plan Details:
The following data management plan has been developed for Professor Periwinkle's research which involves tracking marine wildlife and monitoring ocean conditions through sensor and monitoring equipment and field observations.

MARINE WILDLIFE AND OCEANS DATA MANAGEMENT PLAN

DATA COLLECTION

Data from sensors and monitoring equipment are collected digitally from ROMV, surgically implanted tags in captured and released animals, communication lines that listen for signals from these tags, and static sensor buoys. Sensors produce 300MB of raw data per day. Raw data are converted using proprietary software to NetCDF 4 format which become 500MB/day in uncompressed data. Approximately 500GB of data have been collected to date.

Other data include field notes that document animals captured and tagged and from mark-recapture population estimation experiments and observational studies, approximately 2GB have been collected to date. Citizen scientists share reports downloaded in tab-separated-values (.tsv) format from Professor Periwinkle's website, of which approximately 3GB of data have been collected to date.

Data produced by the buoys and from collaborators are also collected to produce simulation models that track animal populations and movements. These models produce gigabytes-worth of data in zipped comma-separated-values (.csv) format. Visualizations produced using this data are shared with the public on the Oceanviewer.org website.

Data are created and collected in the format specific to each sensor, but converted to NetCDF 4 format. Conversion to NetCDF 4 solves the problem of any specialized or proprietary data formats produced by specific sensors it is an international standard of the Open Geospatial Consortium. The interoperability and accepted standard of the format ensures that datasets are usable and can easily be shared by others, within and outside of Canada.

Field notes currently follow the Darwin Core. However, we would recommend utilizing the GBIF EML profile (version 1.1) metadata standard, as this will allow for data to be moved to the Ocean Biogeographic Information Systems (OBIS) repository. This will allow for linking, sharing, and long-term access to the data as the standard is machine-readable and thus will ensure interoperability. (More on the OBIS repository and why it is recommended can be found in the Data Sharing section of this Plan.)

Data contributed by citizen scientists is captured as *.tsv* files to simplify the storing of data in a tabular structure, and to facilitate the exchange of information between databases, thereby ensuring sharing and interoperability. Data from the simulation models are produced in .csv format which is considered a standard, compact and versatile format that can be opened and edited in various text editors. Both formats will help to ensure data quality, and allow for sharing and long-term access.

We recommend following the guidance contained in the OBIS repository's Integrated Publishing Toolkit when deciding on a folder structure and naming convention for the data. The toolkit offers specific instructions on preparing data to be uploaded to the repository. This includes assistance with mapping datasets to the Darwin Core, and naming conventions for datasets to be published. OBIS has built-in data quality control measures. Any data that do not meet OBIS-accepted standards and cannot easily be corrected are dismissed.

In order to optimize data querying strategies, it is recommended that data be linked by collector device, geographic area of the collection, and time during which collection occurred. These features, and the attributes present in the metadata will allow for sufficient flexibility when executing queries and navigating the contents of specific datasets.

DOCUMENTATION AND METADATA

As mentioned above, in order to comply with OBIS metadata requirements all of the current, future, and past data is required to utilize GBIF EML profile (version 1.1) as a metadata standard. Additional details on EML can be found here: https://knb.ecoinformatics.org/#external//emlparser/docs/index.html and will be included in the documentation provided by our research team. Documentation and additional training will help ensure familiarity with the minimum requirements of OBIS metadata standards. More information can be found here: http://www.iobis.org/manual/eml/

We suggest that the NERC or similar controlled vocabulary be used in conjunction with our recommended metadata standard to ensure searchability and usability of your datasets. The metadata standard and controlled vocabulary outlined in this section will allow the team to search data by organism, geographical features, and environmental conditions and other categories as mentioned in the Data Collection section above.

ETHICS AND LEGAL COMPLIANCE

To ensure ethical compliance, the guidelines established by the institutional Research Ethics Board are to be followed.

Copyright and Intellectual Property Rights (IP/IPR)

As Ocean Biographic Information System (OBIS) is the recommended open access repository for research data, licensing terms will be inherited from repository. OBIS allows for three options: Creative Commons 0 (CC-0), Creative Commons Attribution 4.0 International (CC-BY) and Creative Commons Attribution-NonCommerical 4.0 International (CC-BY-NC). We recommend that CC-0 be used, which is the repository's preferred license. This would place the data into the public domain globally and is legally recognized as a waiving of a researchers' rights to their data. Other users would be able to read, share, modify, enhance, etc. the work freely.

If the preference is to be credited for the work, then we would recommend that CC-BY be used. Users would still be able to read, share, modify, enhance, etc. the data, but they would be required to acknowledge the researcher. CC-BY-NC would allow for the same terms as CC-BY, but the data could not be used commercially. This is the license least recommended by OBIS. For both CC-BY and CC-BY NC, changes made to the data must be noted, so the integrity of the original data is maintained. Creative Commons licenses are reputable and have terms and standards that future researchers will recognize.

STORAGE AND BACKUP

In accordance with best practice, three separate copies of the data should be stored (on different types of media), with at least one copy kept offsite. We further recommend that two copies be kept offsite, which will protect these copies from any unexpected events that may arise in the primary location. While it may be both time-consuming and frustrating to maintain and back up three copies, it is an expected level of due diligence required to reduce the risk of data loss.

As an on-site storage location, we recommend a physical server (see budget) be established. This server will have four hard drive slots so that data expansion is possible and will initially contain a single 1TB drive. The server can be accessed remotely and all team members can upload to this server. We strongly suggest that team access be limited to read, write, modify and execute permissions in order to avoid intentional or accidential deletion of data. We recommend a daily incremental file-based backup with a full backup done weekly using EMC's Avamar software, which installs a backup agent to be pointed to a backup server hosted on Amazon Web Services (AWS) in Montreal.

Cloud infrastructure and storage is a low-cost, encrypted option and also allows for additional storage. We also recommend that all data from the legacy formats (floppy drive, CDs, DVDs, etc.) as well as the external hard drive be transferred to the Cloud backup which can be accessed from the onsite server. As this will be time consuming, we recommend that an RA complete this task. AWS would also serve as a disaster recovery mechanism if data needs to be restored either onsite or to a new location.

Innovation Canada funding requires researchers to use Compute Canada Database (CCDB) for computation. We therefore recommend that an RA also transfer all existing data from the onsite server to CCDB. A properly qualified RA should then write a script that will push incremental changes to CCDB data nightly to ensure data redundancy and grant compliance. It is important to periodically reevaluate backup and storage strategies. We recommend that an RA be responsible for checking the health and effectiveness of the current strategies every six months.

Access and security

To ensure the security measures are in place, data best practice suggests utilizing services that provide two-end encryption. This measure provides a public key and a private key, giving complete control over who has access to your data. Although the present data is not sensitive or personal in nature and thus does not require adherence to Canadian Data Privacy regulations, we recommend that data be stored in Canadian data centres such as AWS-Montreal location, and Compute Canada.

SELECTION AND PRESERVATION

As the current trends in ocean research rely on identifying patterns of animal movement and environmental conditions over time, it is reasonable to keep as much historical data as possible.

This is especially the case for data that cannot easily be recreated according to DataOne best practices.

With this in mind, the collections of formatted NetCDF 4 data created from raw sensor output should be retained over the raw data. For the sake of quality control, raw data will be kept for 6 months prior to being archived.

Data captured from 1998 - 2008 should be standardized into a cohesive format such as NetCDF 4 and archived, with summaries and most critical average values identified for each year and placed in .csv format for repository submission. The next decade (2008 - 2018) should be standardized and sorted by year. Data should then be compressed during backup procedures and placed in the repository, with data from the last 5 years zipped and also stored on an institutional file server, as well as backed up to AWS and sent to Compute Canada.

Long-term preservation plan for the datasets

As noted previously, once the research is complete, we recommend using OBIS as a repository for long-term storage of the research data. OBIS is a large, international, open access repository that is exclusive to marine data collection. It is also funded partly by UNESCO which provides greater assurance in terms of longevity and long-term access. AWS, as noted in the Storage and Backup section, will also provide a long-term storage option from which data may be accessed quickly.

Please note that OBIS does not publish articles. In order to maintain compliance with funding requirements from the Canada First Research Excellence Fund, any article(s) published from this research will need to be made available through another open source provider.

DATA SHARING

We would recommend sharing a link to the OBIS repository. Anyone wanting access to the data can access the repository. Interested parties can upload directly to the repository if they wish to contribute data. The long-term goal would be to set up a file server to facilitate the sharing of data that is currently transmitted through personal connections.

Restrictions on data sharing

It would be advisable to create a document that establishes written acknowledgment of data sharing policies applicable at the end of a student's term as a member of the research team/your lab. Students may be permitted to take copies of data they collected, but originals and ownership over collected data will remain with the Principal Investigator/Professor Periwinkle.

RESPONSIBILITIES AND RESOURCES

Responsibility for data management

Documentation will be created by a designated resource such as a data manager. Best practice and start-up assistance will be provided by F.A.M.E Data Management and transitioned after the end of their responsibilities

Resources required to deliver the data management plan

The following resources will be required:

Information Architecture Resources
Server
Compute Canada
OBIS
Amazon Web Services
Paragon Backup and Recovery

Human Resources
Reaserch Assistant to transfer legacy data
Reaserch Assistant to ensure metadata compliance
Reaserch Assistant to write script for upload to Compute Canada
Data Manager

Financial Resources
Amazon EBS snapshots to Amazon S3 from $.0.055 per GB-Month of data stored (https://aws.amazon.com/ebs/pricing/)
HPE ProLiant ML10 Gen9 - tower - Xeon E3-1225V5 3.3 GHz - 4 GB - 1 TB $ 799
Research Assistant: $23 per hour
Data Manager: $20 per hour