Welcome to the CHONe Data Management Workshop

Please download the following:

- OpenRefine OpenRefine (formerly Google Refine) is a powerful tool for working with messy data: cleaning it; transforming it from one format into another. (Download and extract only, no 'install')

- R is a language and environment for statistical computing and graphics.

- RStudio (the open source 'free' version) is an integrated development environment (IDE) for R. It includes a console, syntax-highlighting editor that supports direct code execution, as well as tools for plotting, history, debugging and workspace management.

- Github Desktop (the app) is a seamless way to contribute to projects on GitHub (the website)

- A Github account Will allow you to host your version controlled project folder (repository) in the cloud for collaboration, sharing (and backup!).

CHONe

CANADIAN HEALTHY OCEANS NETWORK

Welcome

Data Management Workshop

May 1st, 2017

Gatineau, Quebec

Workshop Agenda

CHONe
CANADIAN HEALTHY OCEANS NETWORK

May 1st (6-9pm)

- Introduction to Data Workshop
- Metadata
- Data organization with spreadsheets
- Data cleaning and raw data management

May 2nd (8:30am-5pm)

- Shock and awe with R
- Data Analysis and Visualization in R
- Text analysis
- Data Archiving & Version control

Hacky Hour (6-8pm)

Workshop Agenda

CHONe
CANADIAN HEALTHY OCEANS NETWORK

May 1st (6-9pm)

- Introduction to Data Workshop
- Metadata
- Data organization with spreadsheets
- Data cleaning and raw data management

May 2nd (8:30am-5pm)

- Shock and awe with R ← **Coffee Break 10-10:20am**
- Data Analysis and Visualization in R
- Text analysis ← **Lunch 12-1pm**
- Data Archiving & Version control ← **Coffee Break 3-3:20am**

Hacky Hour (6-8pm)

Network Deliverables

Journal publications

Technical/policy reports

Books

Theses

Educational documents

Technical/analytical frameworks

New statistical/analytical techniques

Maps (pdf, raster, shapefiles..)

Models

Code (R, Python..)

Spreadsheets

Samples

Specimens

Video, Audio and Photos

Partnership meetings

Posters

Presentations

Public forums

Podcasts, blogs and videos

Participation in workshops and conferences and dicussions

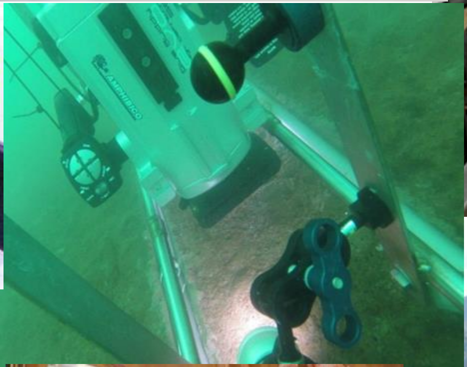
Uptake of research by private enterprise

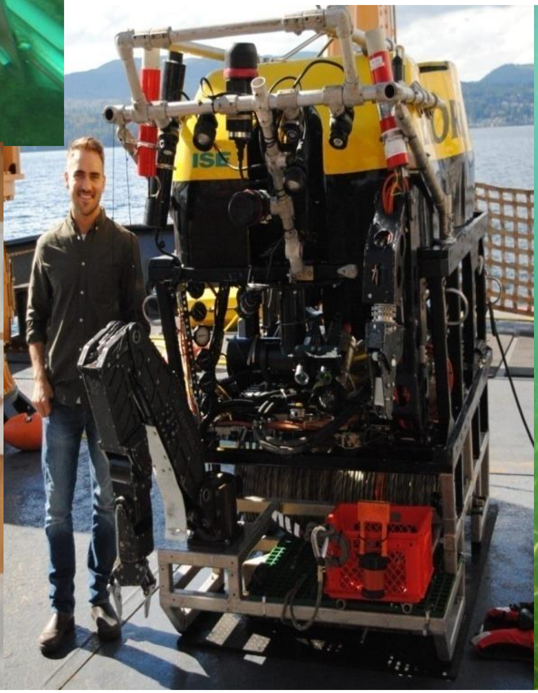
Contribution to management and policy decisions

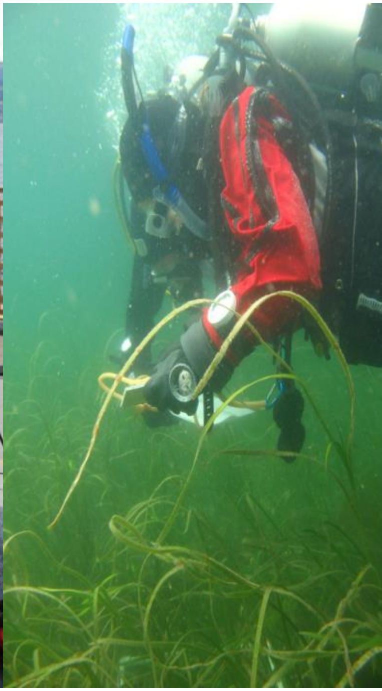
Interviews with radio, tv or newspapers

CHONe Data Efforts...

CHONe
CANADIAN HEALTHY OCEANS NETWORK



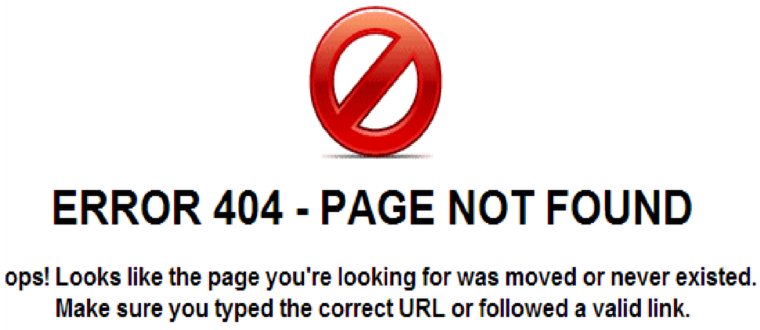




CHONe
CANADIAN HEALTHY OCEANS NETWORK

What can go wrong?

- Boken media device
- Stolen/damaged property
- Hardware/software malfunction
- Building (i.e. fire, flooding, loose power)



ERROR 404 - PAGE NOT FOUND

ops! Looks like the page you're looking for was moved or never existed.
Make sure you typed the correct URL or followed a valid link.

Solution!

Backups:

- Used to take **periodic snapshots** of data in case the current version is destroyed or lost
- Backups are **copies** of files stored for short or near-long-term
- Often performed on a somewhat **frequent schedule**

Archiving:

- Used to **preserve data** for historical reference or potentially during disasters
- Archives are usually the **final version, stored for long-term**, and generally not copied over
- Often performed at the **end** of a project or during major milestones

Data Preservation:

- Includes archiving in addition to processes such as **data rescue**, **data reformatting**, **data conversion**, **metadata**

Solution!

3-2-1 Rule:

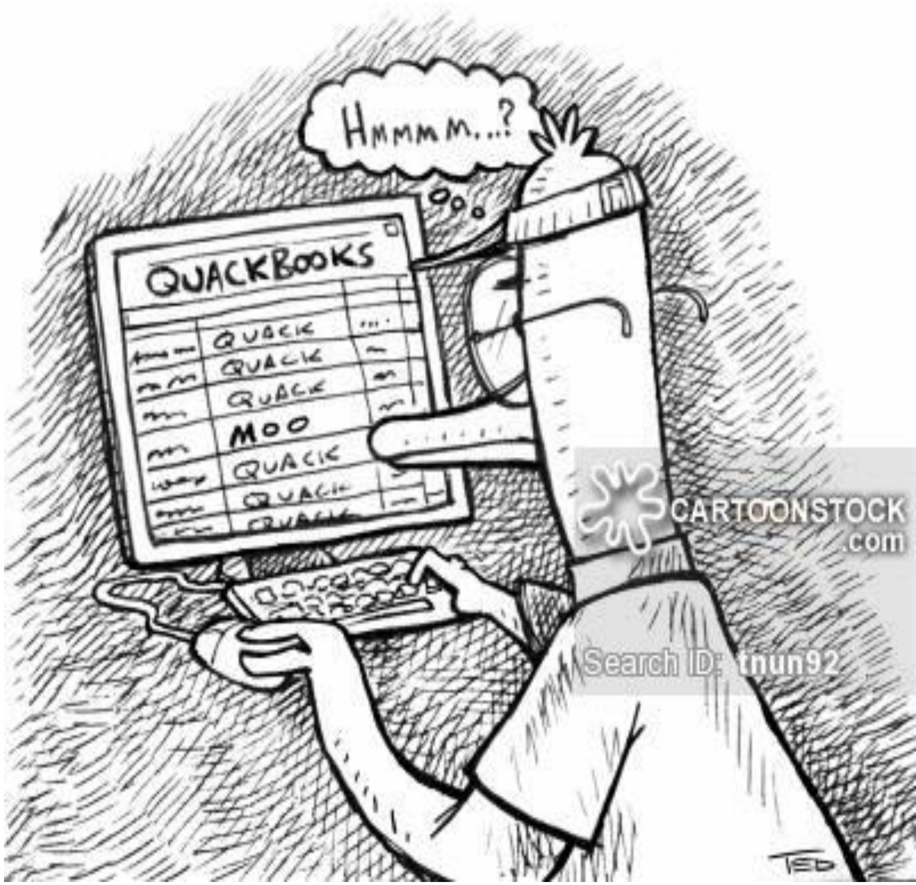
Have at least **three copies of your data**.

Store the copies on **two different media**.

Keep **one backup copy offsite**.

What can go wrong?

- Human error
- Missing data
- Unauthorized changes to data
- Sensor malfunction
- Equipment not calibrated



Solution!

CHONe
CANADIAN HEALTHY OCEANS NETWORK

Quality Assurance and Control: (QA/QC)

- Strategies for **preventing errors** from entering a dataset
- Activities to **ensure quality** of data before collection
- Activities that involve **monitoring** and **maintaining** the quality of data during the study
- Designate **who is responsible for QA/QC** throughout the project

What can go wrong?

- Unspecified acronyms
- Unknown units of measurement
- Methods not described
- Inappropriate use of data
- Not given credit for data use

"Think this is bad? You should see the inside of my head."

Solution!

Metadata:

“Data that provides information about other data”

WHO created the data?

WHAT is the content of the data?

WHEN were the data created?

WHERE is it geographically?

HOW were the data developed?

WHY were the data developed?

Data Management Planning

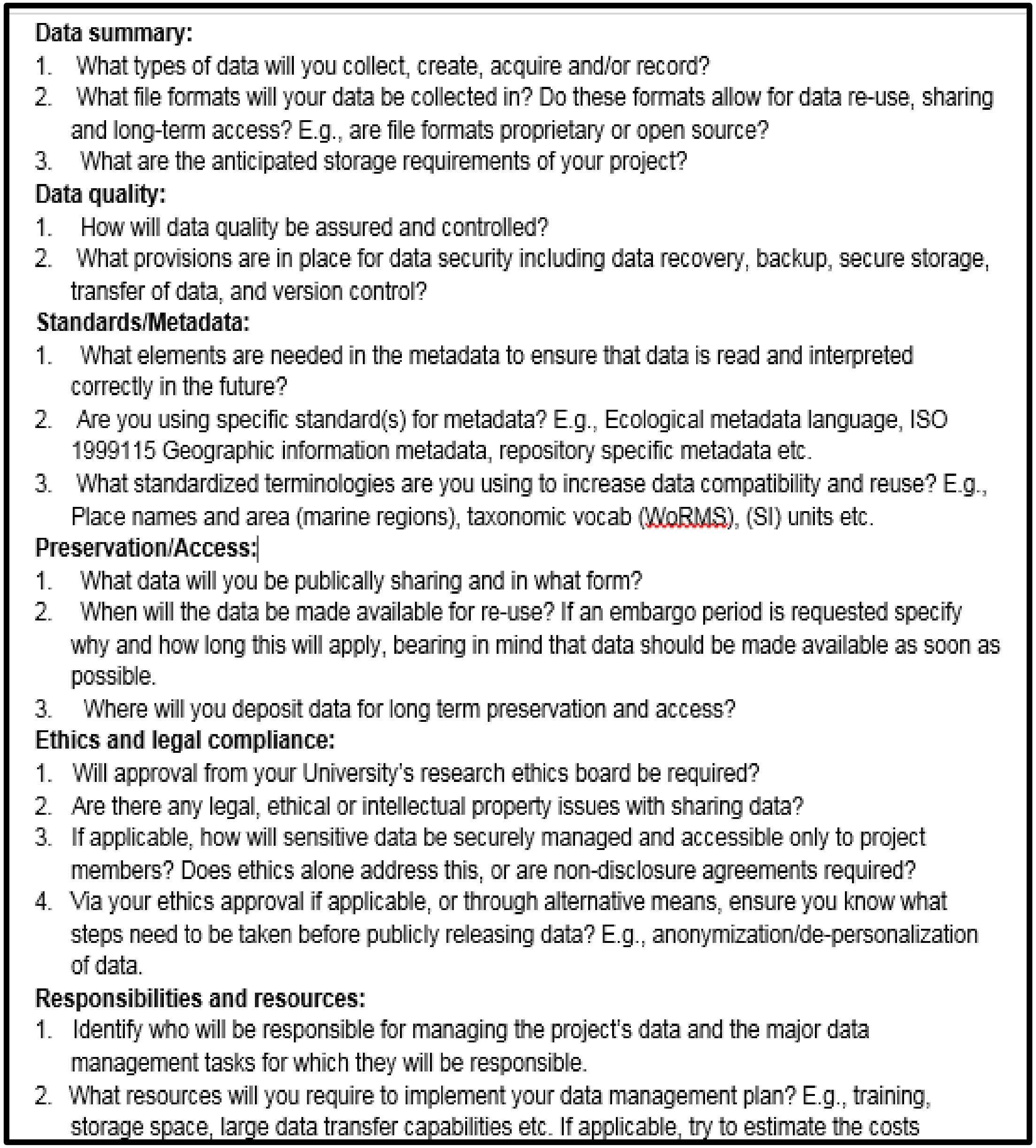
Data Management Questionnaire: a series of questions outlining how you plan to manage your data.

- Structured planning process for each CHONe student and researcher.
- Help identify project requirements early on
  e.g. storage space, software, metadata standards, training workshops…
- Get advice from DM before data collection/analysis
- Fosters network collaboration and data sharing

Data Management Planning

CHONe DMQ:

I. Data Summary
II. Data Quality
III. Standards/Metadata
IV. Preservation/Access
V. Ethics/Legal Compliance
VI. Responsibilities/Resources



Data summary:

1. What types of data will you collect, create, acquire and/or record?
2. What file formats will your data be collected in? Do these formats allow for data re-use, sharing and long-term access? E.g., are file formats proprietary or open source?
3. What are the anticipated storage requirements of your project?

Data quality:

1. How will data quality be assured and controlled?
2. What provisions are in place for data security including data recovery, backup, secure storage, transfer of data, and version control?

Standards/Metadata:

1. What elements are needed in the metadata to ensure that data is read and interpreted correctly in the future?
2. Are you using specific standard(s) for metadata? E.g., Ecological metadata language, ISO 1999115 Geographic information metadata, repository specific metadata etc.
3. What standardized terminologies are you using to increase data compatibility and reuse? E.g., Place names and area (marine regions), taxonomic vocab (WoRMS), (SI) units etc.

Preservation/Access:

1. What data will you be publically sharing and in what form?
2. When will the data be made available for re-use? If an embargo period is requested specify why and how long this will apply, bearing in mind that data should be made available as soon as possible.
3. Where will you deposit data for long term preservation and access?

Ethics and legal compliance:

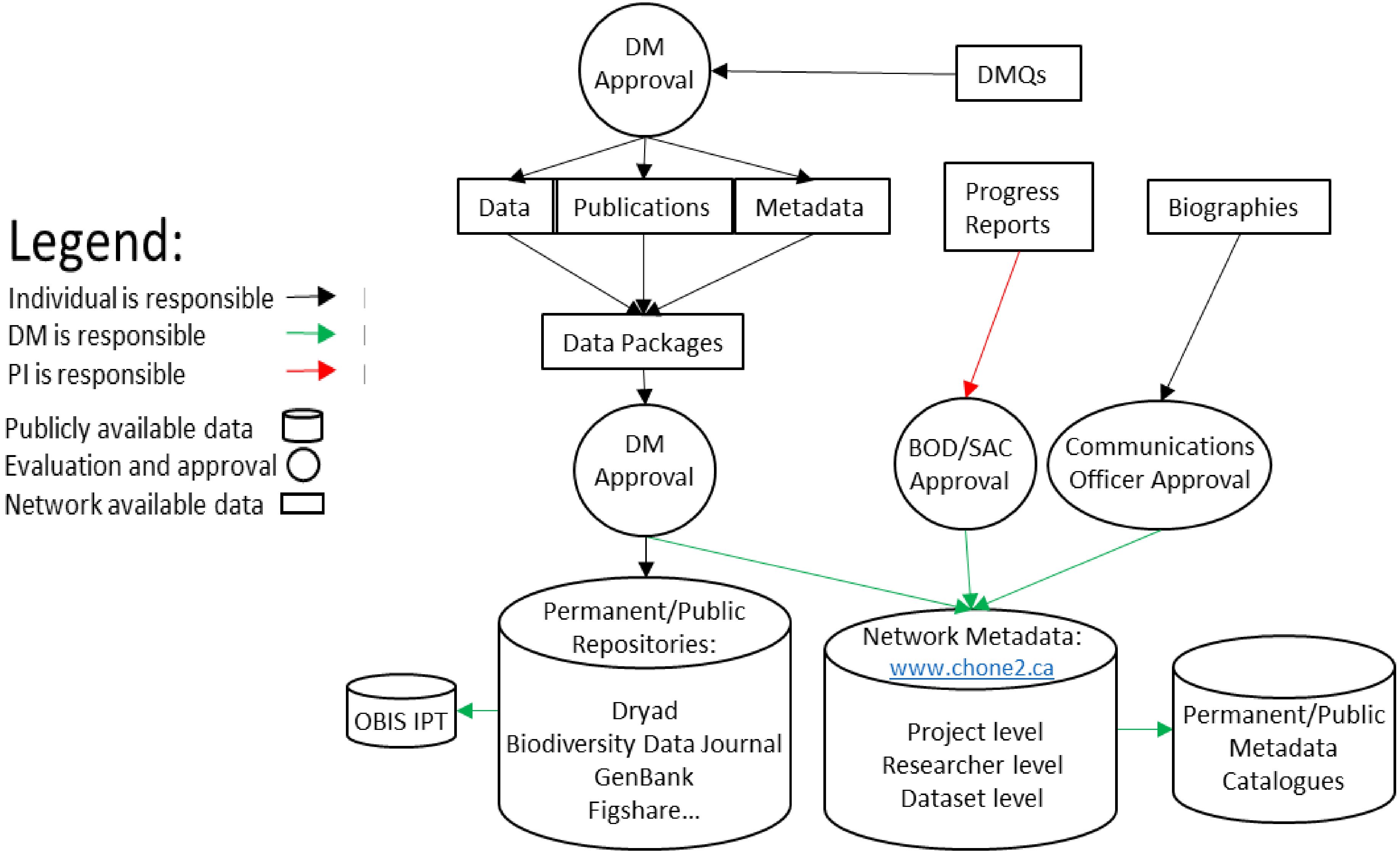
1. Will approval from your University's research ethics board be required?
2. Are there any legal, ethical or intellectual property issues with sharing data?
3. If applicable, how will sensitive data be securely managed and accessible only to project members? Does ethics alone address this, or are non-disclosure agreements required?
4. Via your ethics approval if applicable, or through alternative means, ensure you know what steps need to be taken before publicly releasing data? E.g., anonymization/de-personalization of data.

Responsibilities and resources:

1. Identify who will be responsible for managing the project's data and the major data management tasks for which they will be responsible.
2. What resources will you require to implement your data management plan? E.g., training, storage space, large data transfer capabilities etc. If applicable, try to estimate the costs

CHONe Data Management Plan

CHONe
CANADIAN HEALTHY OCEANS NETWORK



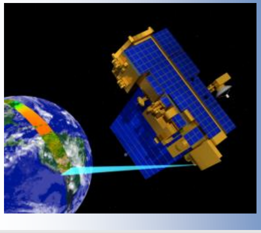
Well-Managed Data Can Result in Re-use, Integration, and New Science

Bird Observations

Land Cover

Meteorology

MODIS – Remote sensing data

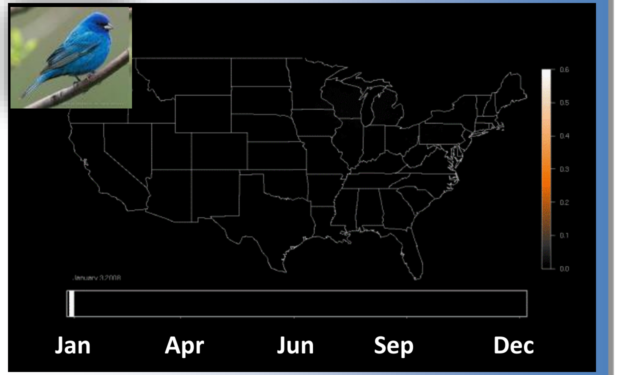


Model: Indigo Bunting locations during the year

$$F(X,s,t) = \frac{1}{n(s,t)} \sum_{i=1}^{m} f_i(X,s,t) I(s,t \in \theta_i)$$

Model results

Occurrence of Indigo Bunting



Potential Uses-

- Examine patterns of migration
- Infer impacts of climate change
- Measure patterns of habitat usage
- Measure population trends

Slide courtesy of DataOne

Why Data Management?

CHONe
CANADIAN HEALTHY OCEANS NETWORK

- Keep yourself **organized**
- Improved **data quality**
- Avoid **data loss**
- Facilitate **data analysis**
- **Reproducibility /accountability** of research
- Get **credit** for your data
- **Promote your research** through data sharing
- Get **hired** after graduation
- **NSERC funding requirements**

Why Data Management?

CHONe
CANADIAN HEALTHY OCEANS NETWORK

- Data as a **public good**
- Data is a **valuable asset** – it is **expensive** and **time consuming** to collect/analyze
- Reduce research duplication
- Speed up **innovation**
- Facilitate **data sharing**
- Foundation to **advance science**

Summary

If data are:

- Well-organized
- Documented
- Preserved
- Accessible
- Verified as to accuracy and validity

Result is:

- High quality data
- Easy to share and re-use in science
- Citation and credibility to the researcher
- Cost-savings to science

Data Management Resources:

Data One: Education Modules

Dalhousie Library: Guide to Research Data Management

UBC Library: Research Data Management

Me (the Data Manager): angela.grant@mun.ca, 1-(709)-864-2298, or on the CHONe Slack page.