GRIIDC Handbook: Data Management Best Practices

Contents

Introduction

The Gulf of Mexico Research Initiative Information & Data Cooperative (GRIIDC) is a leading resource for researchers to manage and share data about the Gulf of Mexico. Proper data management during the course of a project will enable data to be easily shared through GRIIDC or a national data archive. If data is not properly managed, it may be lost or improperly documented, preventing the researcher from sharing and getting credit for work completed.

GRIIDC has compiled a list of 8 data management best practices that can be applied in any discipline from sources listed in the Data Management Resources section of this document (Hook *et al* 2010, Cook *et al* 2001). We have also compiled a list of discipline-specific data management references that can be accessed for additional information. GRIIDC staff are available to provide support and to answer any questions you may have about data management practices.

1. Create and Provide Documentation

GRIIDC requires a description about the context and contents of your dataset. This descriptive information is provided through Dataset Submission. One of the easiest ways to collect the information required to complete your Dataset Submission is to collect the information during the project, rather than waiting to retroactively gather the information at the end of the project.

To help gather this information during the project, you can create a separate file, such as a readme file, or a new Excel worksheet, to track the information at each step of the project. An example of

information that would be included in a readme File is found in Appendix A. This information can then be transferred into GRIIDC Dataset Submission when you are submitting your dataset to GRIIDC.

Important information to document includes:

- The scientific reason data were collected
- What data parameters were collected, including units and format
- What instrument(s) (including models) were used to collect/generate the data
- What platforms were used to collect/generate data (e.g., meteorological stations)
- The name(s) of the data file(s) that make up the dataset
- What instrument was used to collect spatial/geographic information and what spatial resolution the data was collected at (e.g., ± 5m)
- If codes are used in the dataset (e.g., for sampling stations, parameters, etc.), definitions of what each code means
- When and how frequently the data were collected
- How each parameter was measured or produced (methods)
- For each parameter, the units of measure and the format
- For each parameter, the precision and accuracy (if known)
- Data processing that was performed, including error analysis to remove erroneous data
- Standards or calibrations used, if applicable
- Software used to manipulate the data
- Software needed to open up the data file(s), if specialized software is required
- Quality assurance and quality control methods used
- Date when the dataset was last modified
- Pertinent field notes or supplementary files
- Related or ancillary data sets
- Known problems that limit data use

2. Assign Descriptive File and Folder Names

File and folder names should describe the contents of the file or folder, and include enough information to identify the data. File and folder names may contain information such as study title, location, year(s) of study, data type, and version number. Clear, descriptive, and unique file names can help organize your data so that you can find it later and are important when a file is shared with colleagues.

When a user downloads your dataset file from the GRIIDC website, the dataset will be provided with the file name that it was submitted with. GRIIDC does not change dataset file names or names of individual files within datasets. Therefore, GRIIDC recommends using the following guidelines to name your dataset file; if your dataset includes multiple files these guidelines can be followed for all files within your dataset.

- Use descriptive names that indicate what the file/folder contains
- Use short names (less than 50 characters)
- Use simple names that are easy to understand

- Use alphanumeric characters
- Use underscores (_) or dashes (-) rather than spaces
- Avoid special characters such as: \ ‘= /,<>^:;()#*?%,”@!+{}~`[]
- Avoid using internal project codes or acronyms that individuals outside of your laboratory or research group would not understand
- Incorporating the temporal or spatial information when applicable is recommended

You may wish to include the UDI assigned to your dataset in the file name, however, please change the colon (:) to a dash (-). Table 1, below, provides examples of both unsatisfactory and satisfactory dataset file titles.

Table 1: Example Dataset File Names

File Name	Issues
#1.xls	Undescriptive Special Characters
gomri data 12.zip	Undescriptive Spaces
TLF5682.tar	Undescriptive Uses internal project code
Sam37.pdf	Undescriptive Uses internal project name
Dataset publication for the Journal of Examples volume 1 page 32 2012/05/06	Too long Special characters Spaces
PAHConcentrations_ClamTissue2002.csv	None
Smith_CommunitySurveyDataset_2010.txt	None
BacterialProduction_MobileBay_Y1.x999.999-0002.zip	None
Johnson_BubblePlumeVelocity_Lab_Images.tar	None

3. Backup your Data

Unfortunately, data can be lost or deleted accidently; protect against data loss by using a backup solution that 1) runs regularly scheduled or automated backup jobs and 2) stores at least 1 copy at an off-site and trusted location. Ideally your institution or IT Department will have backup services or guidance for your specific computing environment, but GRIIDC provides general suggestions here. Backups should be run regularly, preferably by automated scripts or programs; the frequency depends on how often the data are modified. The off-site location ideally would be another computer or server in a different building or a cloud-based storage system, but could also be an external hard drive. Avoid using USB thumb drives, flash drives, or memory sticks as a main backup solution since these are easily broken and lost.

Once your data is backed up, periodically perform checks to ensure that the backup copies are identical to the original copy of the data; this can be done by using checksums or file differencing programs. Please contact your institutional IT Department for details on how to install and run these programs.

Finally, once you have a backup for your data and periodically afterwards, perform a test of restoring your data from backup and document the steps needed to restore the data. For example, would you need to contact your institutions IT department? If so, who in the IT department would need to be contacted and what would you need to request?

4. Define your Parameters, Units and Formats

Parameters reported in the dataset should have names that describe the contents; these names should be used consistently in different files that make up the dataset. Ideally, these parameters should be the same across datasets created for the same project. Use commonly recognized parameter names and abbreviations. For example, Temp for temperature, Lat for Latitude, and Long for Longitude. Use consistent parameter spelling and capitalization if you are reporting the same parameter in different worksheets or files. For example, always label the parameter Temperature as TEMP; do not call it TEMP, Temp and Temp1 in different worksheets or data files, if they are all reporting the same parameter. If you are using abbreviations make sure you fully describe the abbreviation. Make sure that different parameters have unique names. Some scientific communities have specific controlled vocabularies that are implemented to assist in interpretation of terms by eliminating the use of arbitrary terms that can cause inconsistency and confusion and clarifying and establishing permissible terms. If you are using an established controlled vocabulary, you should document the vocabulary being used and provide this information when submitting your dataset.

Units should be explicitly stated in the data file. For example the parameter temperature may be reported in Celsius, Fahrenheit, or Kelvin. Recording the units that each parameter is measured in will facilitate reuse of the dataset in the future. You should also state the format, if applicable. For example a date can be provided as yy/mm/dd or dd/mm/yy. Define the format for each parameter and use the format consistently whenever the same parameter is reported.

Some datasets will contain null values, missing values, or coded values. Include the notation and definition for null, missing, and coded values. If data are missing, please provide an explanation for the missing data.

If using spreadsheets, one worksheet in the file may list all the parameters and units. A readme file may be created to document this information. By documenting the units during the project, the information will be easily available when submitting data. Data parameters and units are required information for the Dataset Submission in the GRIIDC system.

5. Use Consistent and Stable File Formats

Scientists often use proprietary software to create and analyze data, such as Microsoft Excel, Access, statistical software, and instrument specific software. Unfortunately, storing data using proprietary formats can lead to the data being unreadable in the future. For example, the software could fall out of general use and no longer be supported or available or new updates to the software may make datasets

saved in older versions unreadable. Therefore in addition to saving your file using proprietary software, you should also export and share your files in a non-proprietary file format.

For tabular and other text-based data, CSV (comma-separated values) files are recommended. CSV files are ASCII (American Standard Code for Information Interchange) formatted and are readable by many programs such as Excel or Notepad. To save an Excel file as a CSV, go to "Save As" and select "CSV (Comma Delimited)". Note that you must create a separate .csv file for each worksheet in the Excel file using this method. Analysis, figures, and summary statistics should be provided in separate files from the data. You can then package all the related files together to create a single file using zip or tar and submit the single file as a single dataset to GRIIDC.

For image or raster data files, some non-proprietary file formats available include GeoTIFF/TIFF, HDF5, and ASCII grid. If your data cannot be converted to a non-proprietary file format you should indicate the software program needed to open the data file in the descriptive information provided during Dataset Submission.

Vector data can be stored in shapefiles or GML. For shapefiles please ensure that all the component files are present (can be up to 7 files) and that the projection has been defined. For multi-dimensional data files, NetCDF or HDF5 are stable, widely used formats supported by many open source projects.

6. Use Consistent Data Organization

Tabular data, which is data usually presented in a table format, can usually be organized in one of two ways.

In the first way each observation is a separate row, and each column represents all the parameters for the record. See example below:

Station	Salinity	Temp	Oxygen	Lat	Long
Site1	36	25	4.69	27.5097	-93.3380
Site2	36	25.5	4.69	26.6489	-87.0098
Site3	36	25	4.68	22.6812	-87.7130

Note that each cell or value in a sheet should only include information for one parameter.
If not all parameters are measured consistently and a dataset has a number of missing values, it may make more sense to define the parameter, value, and unit of measure in separate columns. See the example below:

Station	Date	Parameter	Value	Unit
Site1	20140506	Salinity	36	ppt
Site1	20140506	Oxygen	4.69	ml/l
Site2	20140507	Oxygen	4.69	ml/l
Site3	20140507	Salinity	36	ppt

Whatever method is used to organize data should be used consistently, with consistent parameter names, units, and formats. There may be discipline specific standards or best practices for organization

of data available; if so use these standards and best practices and document which standards or practices you have used.

7. Perform Basic Quality Assurance

In addition to the regular quality assurance and control procedures performed regularly by your lab on data, additional quality assurance measures are suggested.

1. Check the file format: make sure that data line up in the correct columns and rows, especially if you have saved a file as a .txt file.
2. Check file organization: make sure that the file contains no missing values; if there are missing values make sure that the reason for these missing values is documented. One suggestion for missing values is to use an extreme value (e.g., -9999) in the cell to indicate that the measurement is missing, rather than the result of an error during data entry or analysis. The purpose of this extreme value should be documented.
3. Check parameter values: look at your dataset for impossible values; for example a negative value when negative values are impossible. Creating plots of your data may help find such values.
4. Check geographic extent data: if latitude and longitude are reported in your data, use a scatter plot or GIS software to map the locations and check to see if there are any errors in coordinates.
5. Check data transfers: when data are being transferred from lab notebooks, data loggers, or other instruments by hand, consider double data entry (entering data twice, and comparing the two datasets and reconciling any differences). When possible, compare summary statistics before and after transferring data.

8. Assign Descriptive Dataset Titles

A dataset title is different from file and folder names within a dataset. Dataset titles are used to cite datasets in publications and are displayed on the GRIIDC website, whereas file and folder names are not visible until a dataset is downloaded. File and folder names are not used to reference datasets in publications.

When naming your dataset in the GRIIDC System we encourage the use of descriptive dataset titles. The title of your dataset should explain what types of data are in your dataset and if applicable should include information about the location and time period that your dataset applies to. It should be understandable to a user unfamiliar with your methods, collection sites, or research platforms. It should not be the same as the title of a publication and does not need to reflect the purpose or results of data collection and analysis. Some examples of good dataset titles are listed here for your reference.

(1) Ecological:
Aerial survey data for the assessment of the distribution of cownose rays (*Rhinoptera bonasus)* in the Eastern Gulf of Mexico, from May to October 2008

(2) Chemical/Molecular Engineering
Image sequences of rising bubble plumes from laboratory study to calculate bubble velocity vector under various gas flow rates and vertical density gradients

(3) Oceanographic
Conductivity, temperature and depth data for 12 northwestern Gulf of Mexico locations, May to July 2012

(4) Model/Numerical Model
Galveston Bay Circulation Study: Stanford Unstructured Nonhydrostatic Terrain following Adaptive Navier-Stokes (SUNTANS) models simulations for 2007-2011

(5) Sociology
Cross-sectional household survey response data to assess health and wellbeing of residents of coastal Louisiana, April 2012

(6) Genetics
Mitochondrial DNA control region sequences (297 base-pairs) from 140 northern Red Snapper (*Lutjanus campechanus*) collected from the Gulf of Mexico, 1998–2001

(7) Biochemistry
Polycyclic Aromatic Hydrocarbon concentrations in liver and muscle tissue from barracuda, escolar and common dolphin fish, northeastern Gulf of Mexico, 2011-2012

Data Management Resources

GRIIDC has compiled the following list of other data management practice resources for your reference. This list is not exhaustive and we encourage you to contact GRIIDC if with suggestions for additional resources to include in this section.

General Data Management Resources

DMPTool. Data Management General Guidance. University of California Digital Library. Accessed online October 3, 2014. Available at: https://dmptool.org/dm_guidance.

Schmitt, C.P. and M. Burchinal. 2011. Data Management Practices for Collaborative Research. *Frontiers in Psychiatry.* 47: 1-8. doi: 10.3389/fpsyt.2011.00047. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3143734/pdf/fpsyt-02-00047.pdf.

University of Oregon Libraries. Research Data Management Best Practices. Accessed online October 3, 2014. Available at: https://library.uoregon.edu/datamanagement/guidelines.html

Discipline Specific Data Management Best Practices Resources

Ecology

Borer, E.T., Seabloom, E.W., Jones,M.B., and M. Schildhauer. 2009. Some Simple Guidelines for Effective Data Management. *Bulletin of the Ecological Society of America* 90:205-214. doi: 10.1890/0012-9623-90.2.205

Cook, R.B., Olson, R.J., Kanciruk, P. and L.A. Hook. 2001. *Best Practices for Preparing Ecological Data Sets to Share and Archive.* Bulleting of the Ecological Society of America 82:138-141. Available at: http://www.esajournals.org/doi/pdf/10.1890/0012-9623%282001%29082%5B0136%3AC%5D2.0.CO%3B2

Whitlock, M.C. 2011. Data archiving in ecology and evolution: best practices. *Trends in Ecology and Evolution* 26:61-65. doi: 10.1016/j.tree.2010.11.006.

Economic and Social Science

Van den Eynden, V., Corti, L, Woollard, M., Bishop, L. and L. Horton. 2011. Managing and Sharing Data: Best Practices for Researchers Third Edition. UK Data Archive. University of Essex. 35pp. Available at: http://www.data-archive.ac.uk/media/2894/managingsharing.pdf.

Environmental

DataONE All Best Practices. Downloaded on October 3, 2014. Available at: https://www.dataone.org/all-best-practices-download-pdf

Hook, Les A., Suresh K. Santhana Vannan, Tammy W. Beaty, Robert B. Cook, and Bruce E. Wilson. 2010. Best Practices for Preparing Environmental Data Sets to Share and Archive. Available online (http://daac.ornl.gov/PI/BestPractices-2010.pdf) from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/BestPractices-2010.

Strasser, C., Cook, R., Michener, W. and A. Budden. Primer on Data Management: What you always wanted to know but were afraid to ask. DataOne: Albuquerque, NM. Available at: https://www.dataone.org/sites/all/documents/DataONE_BP_Primer_020212.pdf

Fisheries

Kolb, T., Blukacz-Richards, E.A., Muir, A., Claramunt, R.M., Koops, M.A., Taylor, W.W., Sutton, T.M., Arts, M.A., and E. Bissel. 2013. How to Manage Data to Enhance Their Potential for Synthesis, Preservation, Sharing, and Reuse – A Great Lakes Case Study. *Fisheries* 38:52-64. doi: 10.1080/03632415.2013.757975.

Marine Biogeochemistry

Pollard, R.T., Moncoiffé G., and T.D., O'Brien. 2011. The IMBER Data Management Cookbook – A Project Guide to good Data practices. IMBER Report No. 3, IPO Secretariat, Plouzané, France. 16 pp. Available at: http://www.imber.info/index.php/content/download/1158/5654/file/IMBER%20Data%20Management%20cookbook.pdf

Oceanography

BCO-DMO Data Management Guidelines Manual, v1.0. 2008. Available online at: http://www.bco-dmo.org/files/bcodmo/BCO-DMO_best_prac_v1d2.pdf

Social Science

Burchinal, M and E. Neebe. 2006. Data Management: recommended practices. *Monographs of the Society for Research in Child Development* 71:9-23. doi: 10.1111/j.1540-5834.2006.00402.x . Available at http://onlinelibrary.wiley.com/doi/10.1111/j.1540-5834.2006.00402.x/pdf

Inter-university Consortium for Political and Social Research (ICPSR). 2012. Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle (5th ed.). Ann Arbor, MI. Available at: http://www.icpsr.umich.edu/files/ICPSR/access/dataprep.pdf

Appendix A: Example ReadMe File Information

Note: A read me file should be saved as a .txt file and can be included in a dataset by zipping it together with the other files in your dataset.

```
================================================================================
Sensor heights (in meters above mean water level)
================================================================================

        Gill wind/sound speed sensor: 11.4
                  LI-COR IRGA sensor: 10.9  (CO2/H20 concentrations)
  Rotronic RH/air temperature sensor: 10.7
            Setra air pressure sensor:  9.9

================================================================================
GLAD_ship_met_1.dat and GLAD_ship_met_2.dat
================================================================================
Records averaged to 1-second intervals
================================================================================

Column 1:  date (yyyy-mm-dd)
Column 2:  time (HH:MM:SS)
Column 3:  water vapor concentration (mmol/m^3)
Column 4:  CO2 concentration (mmol/m^3)
Column 5:  relative humidity (%)
Column 6:  air temperature (Rotronic sensor) (degrees C)
Column 7:  air pressure (hectopascals, hPa)

================================================================================
GLAD_ship_GILL_1.dat and GLAD_ship_GILL_2.dat
================================================================================
Records averaged to 1-second intervals
================================================================================

Column 1:  date (yyyy-mm-dd)
Column 2:  time (HH:MM:SS)
Column 3:  U, boward wind speed (m/s)
Column 4:  V, aport wind speed (m/s)
Column 5:  W, upward wind speed (m/s)
Column 6:  C, speed of sound (m/s)

================================================================================
GLAD_ship_positions.dat:  Ship GPS positions
================================================================================
Records interpolated to 1-second intervals
================================================================================

Column 1:  date (yyyy-mm-dd)
Column 2:  time (HH:MM:SS)
Column 3:  longitude (decimal degrees)
Column 4:  latitude  (decimal degrees)
```