BEST PRACTICES FOR SHARING AND ARCHIVING DATASETS

October 2011

Modified by Josée Michaud and Julie Friddell with the permission of L.A. Hook, "*Best Practices for Preparing Environmental Datasets to Share and Archive*," updated by L.A. Hook, T.W. Beaty, S. Santhana-Vannan, L. Baskaran, and R.B. Cook, June 2007. Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. Environmental Sciences Division, Oak Ridge National Laboratory. Previously entitled "*Best Practices for Preparing Ecological and Ground-Based Datasets to Share and Archive*," Cook *et al.*, 2001. (http://daac.ornl.gov/PI/bestprac.html)

Table of Contents

Introduction

The following guidelines have been prepared to facilitate effective data management practices. Data collectors and data managers should follow these guidelines to improve the usability of datasets. This guidance is provided for a variety of data collections and archiving activities arising from the Polar Data Catalogue (PDC, www.polardata.ca), ArcticNet, and the International Polar Year (IPY). These guidelines are formulated for datasets in the fields of physics, atmospheric science, oceanography, ecology, contaminants, human health, and social sciences but will certainly be useful for other types of data collections.

In preparing datasets for contribution to the PDC, researchers should determine how finely or coarsely to divide their data, considering primarily the coherence and future usability of the data in the dataset. The overarching goal of archiving and preserving data is to make it easily accessible and reusable for people unfamiliar with the project.

Assembled here are the most important practices that researchers could implement to make their datasets ready and easy to share with other researchers and interested parties.

In summary, guidelines outlined for these practices are:

1. **Providing Metadata**
2. **Assigning Descriptive Metadata Titles and Data File Names**
3. **Using Consistent and Stable File Formats for Tabular and Image Data**
4. **Defining the Contents of Data Files**
5. **Using Consistent Data Organization**
6. **Performing Basic Quality Assurance**
7. **Providing Dataset Documentation**
8. **Citing a Dataset**

Guidelines

1. Providing Metadata

Metadata is the description of a dataset. Metadata provide the what, where, and when of data, by whom it was collected, and its current location. Metadata should be written for a user twenty years into the future who is unfamiliar with the project, sites, methods, or observations. Such records facilitate the understanding, use, and management of data. Metadata also serve to facilitate networking, collaboration, and data access. For the Polar Data Catalogue, metadata can be accessed at http://www.polardata.ca/. PDC metadata are published in Federal Geographic Data Committee format under metadata standard FGDC Content Standards for Digital Geospatial Metadata, version FGDC-STD-001-1998 (http://www.fgdc.gov/metadata/geospatial-metadata-standards#csdgm). To improve interoperability with other data centres and archives, PDC metadata will be converted to the North American Profile of the ISO 19115:2003 international metadata standard (http://www.fgdc.gov/metadata/geospatial-metadata-standards#nap).

The title of the metadata should be representative of the accompanying dataset (see Section 2). The following information is required in the PDC metadata record:

- ✓ Title of the data
- ✓ How data should be cited (see section 8)
- ✓ Study site
- ✓ Purpose: scientific reason why data were collected
- ✓ Abstract:
  - What data were collected
  - What instruments and source(s) were used
  - When, where, and how frequently the data were collected
  - How parameters were measured or produced (methods), units of measure, precision and accuracy if known, and the relationship to other data in the dataset, if appropriate
- ✓ Data originator, Principal Investigator, and other responsible parties and contact information (e-mail or Web address, if appropriate)
- ✓ Research program
- ✓ Status of data collection – "Complete," "In Progress," or "Planned" if data not available yet
- ✓ Maintenance and update frequency of data
- ✓ Geographic coordinates
- ✓ Time period
- ✓ Keywords
- ✓ Access restrictions – "Public" or "Limited"

2. Assigning Descriptive Metadata Titles and Data File Names

2.1. Metadata

Datasets will be accessed many years in the future by people who will be unfamiliar with the details of the project. Metadata titles therefore need to be as descriptive as possible and should include the time period and location, if applicable. The title should be included in the header rows of the corresponding data file(s) and in companion documents. Titles such as "Hudson Bay data" or "Respiration Data" are examples of titles to avoid. The following are good examples of Metadata titles:

- "Global Warming and Arctic Marine Mammals (GWAMM) - community-based monitoring of narwhal whales near Repulse Bay, Nunavut, 2007-2009"
- "Negotiating Change: Community-based Mental Health and Addictions Practice in the Northwest Territories"

2.2. Data Files

Data file names may contain information such as project acronym, study title, location, investigator, year(s) of study, data type, version number, and file type. It is recommended to include the file name in the header rows of the data file itself. Clear, descriptive, and unique file names are important when data files are combined in a

directory, FTP site, or website which contains data files of many different investigators. General file names such as "mydata.dat" or "2009.txt" should be avoided.

A given dataset might contain only one data file or many thousands of data files. For complex datasets with many files, the directory structure for the files should be logically arranged. The directory names should contain the type of data and other appropriate information such as date range, location, and instruments used.

The following guidelines for naming data files should be respected as much as possible:

- ✓ File names must be easily managed by various data systems and thus should contain only numbers, letters, dashes, and underscores. Spaces and special characters are not allowed. Lower-case names are preferred.
- ✓ For practical reasons of usability in the Polar Data Catalogue, file names must not be more than 150 characters in length.
- ✓ Directory structures and names should be designed with the same logic.
- ✓ Data file creation date or version number should be included in the title as it enables one to quickly determine which data are being used if an update to the data file is released (e.g., *_v1.csv, *_r1.csv, or *_20070227.csv).
- ✓ File Type or Extensions *.txt and *.csv are generally preferred for tabular data but *.xls is acceptable. Avoid *.xlsx as it is not backward-compatible.
- ✓ Readily-accessible file types such as *.zip, *.gz, or *.tar are the most appropriate if files need to be compressed.

Upon submission of the data file, the CCIN Reference Number of the metadata record will be added at the beginning of the file name to allow cross-referencing of the dataset with the metadata record.

2.3. Example Metadata Titles, Directory Structure, and Data File Names to be submitted to the PDC

Metadata record: HPLC Pigment Analysis of the Phytoplankton Community in Franklin Bay.
Data file: CASES_HPLC_Franklin_20090914.xls

Metadata: Variability of berry productivity in the context of climate change in Nunavut, linking local and traditional ecological knowledge.
Data file: AN_berry_TK_NU20090822.txt

Metadata: ArcticNet 0603-Beaufort Sea CTD data.
Data directory name: AN_200603_CTD_Beaufort_ 20090914
Data file names: 0603_001.int;
0603_002.int;
0603_003.int
Etc.

3. Using Consistent and Stable File Formats for Tabular and Image Data

3.1. Tabular Data

It is important to choose a consistent format that can be read long into the future and is independent of changes in applications. Non-proprietary formats such as TEXT or ASCII or other readily accessible (i.e., not requiring restricted-use or costly software) formats will have the longest lifespan and thus will allow for the longest usefulness of archived data. Microsoft Word and Excel are widely used and acceptable but not preferred; Word documents and Excel spreadsheets can easily be converted into text (*.txt) or Adobe pdf (portable document format) files.

3.1.1. Guidelines for tabular data format:

- ✓ Same format should be used throughout the file.
- ✓ Consistent file format should apply to all data files belonging to the same project.
- ✓ Figures and analyses should be reported in companion documentation (not in the data file).
- ✓ Header rows should be inserted at the top of the file. The first row should contain descriptors that link the data file to the dataset and to the metadata. For example, provide the data file name, dataset/metadata title, author, today's date, date the data within the file were last modified, and companion file names.
- ✓ Column headings should describe the content of each column.
- ✓ Column headings should be constructed for easy importing by various data systems.
- ✓ Headings should contain only numbers, letters, and underscores - no spaces or special characters. Lowercase letters are preferred.

3.1.2. Guidelines for text file format:

- ✓ Column headings and parameter fields should be delimited using commas, tabs, or semicolons.
- ✓ Delimiters that also occur in the data fields should be avoided. If this cannot be avoided, enclose data fields that also contain a delimiter in single or double quotes.
- ✓ A semicolon should be used as column delimiter if the data fields use the comma as the decimal separator (rather than the period).

3.2. Image (Raster) Data

Some field researchers may generate image (raster) datasets. Below are some recommendations for archiving these types of data files. Suggested non-proprietary or readily-accessible file formats:

- ✓ GeoTIFF/TIF (*.tiff, *.tif)
- ✓ ASCII Grid (*.asc, *.txt, *.flt)
- ✓ Binary image files (BSQ/BIL/BIP) (*.bsq, *.bil, *.bip)
- ✓ Net-CDF (*.nc)
- ✓ HDF (-EOS)

If the above formats are not suitable, further information on the documentation requirements for contributing data in non-proprietary, public domain formats can be found in **Appendix A**.

3.3. Image (Vector) Data

Below are suggested vector file formats. These are mostly proprietary data formats; Software package, version, vendor, and native platform should be documented.

- ✓ ARCVIEW software – *.shp, *.sbx, *.sbn, *.prj, and *.dbf files that contain the basic components of an ARCVIEW shape file (http://www.esri.com/)
- ✓ ENVI - *.evf (ENVI vector file) (www.rsinc.com/whoweare/index.asp)
- ✓ ESRI Arc/Info export file (*.e00) (www.esri.com)

Vectors must be properly geo-referenced and the geometry type (point, line, polygon, multipoint etc.) specified.

3.4. Photos and Videos

Common formats such as JPG, PNG, mpeg, wmv, and avi are preferred.

3.5. Proprietary Software Data Formats

Data that are provided in a proprietary software format must include documentation of the software specifications (i.e., software package, version, vendor, and native platform). The PDC data managers may use this information to convert to a non-proprietary format for the archive. MS Office products are proprietary formats which may be subject to conversion problems in the future. These file types should be converted, prior to upload to the PDC, to non-proprietary formats when possible.

For geographic data, all file types that constitute the complete geographic data format documentation must be provided for the specific software package. For example:

- ✓ idrisi software images - provide the *.rdc and the *.rst files (www.clarklabs.org)
- ✓ IMAGINE software images - provide *.img and *.rrd files (http://gis.leica-geosystems.com)
- ✓ ENVI images - provide ENVI *.hdr file (http://www.ittvis.com)

To find out what program can be used to open any file format:
http://filext.com/index.php

4. Defining the Contents of Data Files

Users must be provided adequate information to fully understand the content of datasets, including parameter names, units of measure, formats, and definitions of coded values. It is important to provide the English language translation of any data values and descriptors that are in another language (e.g., coded fields, variable classes, geographic information system (GIS) coverage attributes). This information can be provided in header rows of data files, in companion documents, or in the metadata records.

4.1. Parameter Names

The parameter names reported in the dataset must clearly define their content. The documentation should also contain full parameter descriptions. Commonly accepted parameter names should be favoured (e.g. "Temp" for temperature, "Precip" for precipitation, "Lat" for latitude, etc.). It is important to use consistent capitalization (not temp, Temp, and TEMP) and use only letters, numerals, and underscores in the parameter names (no spaces or special characters).

4.2. Units

The units of all reported parameters need to be explicitly stated in the data file and in the documentation so users understand what is reported. The International System of Units (SI) is preferred. If a shorthand notation is reported in the data file, complete units should be provided in the documentation.

4.3. Formats

Parameter formats should be explained in the dataset documentation and be consistent throughout a dataset. Consistent formats are particularly important for dates, times, and spatial coordinates. It is also important to explicitly define the format of numeric parameters with significant digits so significant figures are not lost or gained during data transformations by new users.

The following formats for common parameters are recommended:

- ✓ Dates: yyyy-mm-dd or yyyymmdd (e.g., January 2, 1997 is 19970102).
- ✓ Time: 24-hour notation (e.g., 13:30 hrs or 1330 hrs instead of 1:30 p.m.). Both local time and Coordinated Universal Time (UTC) should be reported for both the begin and end times of the data collection.
- ✓ Spatial Coordinates: decimal degrees to at least 4 (preferably 5 or 6) digits past the decimal point. By convention, South latitude and West longitude are negative. All location information in a file should use the same coordinate system (e.g., WGRS84, NAD83), and the coordinate system should be clearly identified.
- ✓ Elevation: meters. Information on the vertical datum used should be provided (e.g., North American Vertical Datum 1988 (NAVD 1988)).

4.4. Coded Fields

Coded fields (e.g. postal codes) offer standardized lists of predefined values for data management and are more efficient for storage and retrieval of data than free text fields. Custom coded fields with defined values, such as sampling site designations, may be created and consistently used across multiple data files.

Data flag or qualifying values are coded fields commonly used in environmental data to indicate quality considerations, reasons for missing values, replicated samples, etc. Codes should not be parameter specific but should be consistent across parameters and data files. Definitions of flag codes should be included in the accompanying dataset documentation.

Example of Data Quality Flag values:

Flag Value	Description
V0	Valid value
V3	Valid interpolated value
M1	Missing value because no value is available
M2	Missing value because invalidated by data originator
…	…

Missing values are best represented by a coded field. A specified extreme value not likely to be confused with a measured value (e.g., -9999) is preferable. Consistent notation for each missing value in the data file should be used to represent missing values in numeric fields and should not contain character codes. For character fields, it may be appropriate to use "Not applicable" or "None" depending upon the organization of the data file.

For further information, an example of effective parameter documentation is provided in **Appendix B**.

5. Using Consistent Data Organization

Data within a file can be arranged in numerous ways. In the most common arrangement, each row in a file represents a complete record, and the columns represent all the parameters that make up the record. This arrangement is similar to a spreadsheet or a matrix, as in the following example:

Cruise_Number:	2002002					
Cruise_Name:	CASES	202				
Original_Filename:	CTD_2002002_001_1_DN.ODF					
Station:	1					
Cast_Number:	1					
Start_Date_Time	[UTC]:	22/09/02	09:48.0			
Initial_Latitude	[deg]:	70.3498				
Initial_Longitude	[deg]:	-123.8988				
Sounding	[m]:	318				
Min_Depth	[m]:	1.54				
Max_Depth	[m]:	308.93				
Depth (m);Temp (Celcius); Sal (PSU); Dens (kg/m3); pH (no units); O2 (Dissolved, ml/l); PAR (µEinsteins/m2/sec)						
Depth (m)	**Temp**	**Sal**	**Dens**	**pH**	**O2**	**Par**
2	0.88	23.652	18.94	8.14	7.922	2.481
3	0.78	23.887	19.14	8.138	7.899	1.93
4	0.689	24.15	19.36	8.141	7.855	1.532
5	0.571	24.555	19.69	8.139	7.781	1.224
6	0.278	25.348	20.34	8.13	7.762	0.968
7	0.005	26.132	20.99	8.123	7.79	0.745

8	-0.484	27.405	22.03	8.113	7.845	0.588
…	…	…	…	…	…	…

[From Gratton et al., 2002, CASES 0202-Beaufort Sea CTD data. Metadata available at http://www.polardata.ca, CCIN reference number 513.]

Other datasets, especially qualitative data, may require different arrangements. In such cases, clear explanation on how the data are organized should be provided. The most important consideration is to ensure that data organization is consistent throughout all the files of a dataset. See **Appendix C** for an alternative example of data organization.

5.1. Keeping Similar Information Together

An important issue with data organization is the number of records in each file. In general, it is preferable to keep a set of similar measurements together (e.g., same investigator, methods, time basis, and instruments) in one data file and not break up data into many small files (e.g., by month or by site when there are several months or sites). There is however an upper size limit to files and large files do become unmanageable for some applications. Large tabular data files may need to be logically broken into smaller files (e.g. water column profiles divided into one file per station).

5.2. Organization by Data Type

For each data file, it is important to keep similar data organization, parameter formats, and common site names, so that users understand the interrelationships between data files. Data types collected on different time bases (e.g., per hour, per day, per year) might be handled more efficiently in separate files. Alternatively, if relatively few observations are made at a site for a suite of parameters, then all data could be placed in one file.

6. Performing Basic Quality Assurance

In addition to scientific quality assurance (QA) that the researchers must perform on data prior to contributing files to the data repository, basic data QA should be performed on data files. As stated in the PDC terms of use, the database and all its content are provided "as is" without quality warranties of any kind Data managers for the PDC can verify file format and documentation but will not be performing QA on data sets. The following items complement the tabular and image file preparation guidance provided in Section 3 and should be verified by the researcher prior to upload to the PDC.

6.1. Tabular Data

- ✓ File format must be consistent and data delimited/line up in the proper column.
- ✓ Missing values for key parameters (e.g. sample identifier, station, time, date, geographic coordinates) must be verified.
- ✓ Documentation should be reviewed to ensure that descriptions accurately reflect the data file names, format, and content. Included example data records should be from the latest version of the data file.
- ✓ Content of measured or derived values should be verified.
- ✓ Parameters should be scanned for impossible values.
- ✓ Geographic coordinates should be checked for errors.

6.2. Image Vector and Raster Data

For GIS image/vector files, ensure that the projection parameters have been accurately given. Additional information such as data type, scale, corner coordinates, missing data value, size of image, number of bands, and endian type should be checked for accuracy.

7. Providing Dataset Documentation

Metadata and accompanying documents need to be written for a user who is unfamiliar with the project, sites, methods, or observations. Written documentation should be in TEXT format. If figures, maps, equations, or pictures need to be included, document formats such as .pdf (Portable Document Format) or .html (hypertext markup language) can be used. Images, figures, and pictures may be included as individual gif (graphics interchange format) or jpg (Joint Photographic Experts Group) files.

⇒ Thorough documentation is critical to the long-term usefulness of archived datasets. A template form with instructions to facilitate complete documentation of datasets has been provided in **Appendix D**. *Data contributors are encouraged to use this README template when completing the dataset documentation.*

8. Citing a Dataset

The following guidelines were drawn from the International Polar Year Data and Information Service (http://ipydis.org/data/citations.html). In general, dataset citation should follow the author-date system, as seen for books. The citation should include the following information, if available:

Author(s)
Publication date
Title
Editor or compiler
Publication place
Publisher
Distributor
Distribution location
Access date
Data within a larger work

8.1. Examples

8.1.1. Author-year

Gratton, Y., 2002, *CASES 0202-Beaufort Sea CTD data*. Canadian Cryospheric Information Network (CCIN). Metadata accessed 2009-10-13 at www.polardata.ca, CCIN reference number 513.

Oberbauer, S., 2000, *Ecosystem carbon fluxes, Toolik Lake, Alaska 1995*. Boulder, Colorado USA: National Snow and Ice Data Center. Dataset accessed 2010-11-16 at http://nsidc.org/data/arcss006.html.

Canadian Snow Data CD by Province, 2003, Meteorological Service of Canada, Environment Canada, Ottawa. CD-ROM.

8.1.2. Versions

Hall, Dorothy K., George A. Riggs, and Vincent V. Salomonson. 2007, updated daily. *MODIS/Aqua Sea Ice Extent 5-Min L2 Swath 1km V005*, Oct. 2007–Apr. 2008. Boulder, Colorado USA: National Snow and Ice Data Center. Dataset accessed 2010-11-16 at http://nsidc.org/data/myd29v5.html.

Cavalieri, D., C. Parkinson, P. Gloersen, and H.J. Zwally. 1996, updated 2006*, Sea ice concentrations from Nimbus-7 SMMR and DMSP SSM/I passive microwave data*, March 2002–Sept. 2003. Boulder, Colorado USA: National Snow and Ice Data Center. Dataset accessed 2010-11-16 at http://nsidc.org/data/nsidc-0051.html.

8.1.3. Editor or Compiler

Armstrong, R., J. Francis, J. Key, J. Maslanik, T. Scambos, and A. Schweiger, 1998, *Polar Pathfinder sampler: Combined AVHRR, SMMR-SSM/I, and TOVS time series and full-resolution samples.* Compiled by S. Khalsa. Boulder, CO, USA: National Snow and Ice Data Center. CD-ROM.

Environmental Working Group, 2000, *Environmental Working Group: Joint U.S.-Russian Arctic sea ice atlas*. Ann Arbor, MI: Environmental Research Institute of Michigan; distributed by the National Snow and Ice Data Center. CD-ROM.

8.1.4. Digital Object Identifier (doi)

König-Langlo, Gert and Hatwig Gernandt, 2006, *Compilation of radiosonde data from the Antarctic Georg-Forster station of the German Democratic Republic from 1985 to 1992.* Bremerhaven, Germany: Alfred Wegener Institute for Polar and Marine Research. Dataset accessed 2008-05-22. doi:10.1594/PANGAEA.547983

Acknowledgements

We thank Les Hook and collaborators for graciously giving permission to use their document for preparing this PDC Best Practices manual. We also thank Julie Driver for the initial work on this document. We are also grateful for the feedback provided by our colleagues with the IPY Data Assembly Centre Network.

Appendix A - Requirements for Alternate Formats for Image (Raster) Data

If the raster formats listed in Section 3.2. are not suitable, non-proprietary public domain data formats can be used. Thorough documentation on the format should be provided using the following guidelines:

- ✓ Consistent naming conventions should be used. Image product name and creation date of the product should be provided as part of the file name.
- ✓ All filenames and file sizes should be listed in the dataset documentation for user verification purposes.
- ✓ Detailed information about the data file format should be provided.
- ✓ Projection information: geographic coordinate system, datum, corner/centre pixel coordinates, etc. If appropriate, projection pre-set or a companion header file with projection information should be included. Example header files: ENVI, *.hdr file; TIF world file, *.tfw; ESRI projection file, *.prj.
- ✓ Indicate whether the corner pixel coordinates lie within the center of the pixel or at one of the edges.
- ✓ Latitude and longitude of the first and the last data values should be provided when a raster dataset is provided as an ASCII/Binary file.
- ✓ Consistent NODATA values are preferred (-9999 or 0).
- ✓ Complete spatial coverage of the data should be given: Upper Left, Lower Right, Lower Left, and Upper Right coordinates. Also, coverage information should be provided in both the native coordinate system and in Geographic Latitude and Longitude.
- ✓ Pictures of binary image files (for example, *.jpg, *.png, *.gif, *.bmp, .tif, or .tiff pictures of geographic images) should be included so user can verify that the binary images were read in correctly.
- ✓ Generic file extensions (e.g., *.dat or *.img) should be avoided to prevent confusion on their origin.
- ✓ Information on software package and version used to create the data file(s) should be given. If the data files were created with custom code, a software program should be provided in the documentation to enable the user to read the files (e.g., FORTRAN, C code, etc).

Appendix B - Example of Effective Parameter Documentation

The following describes the parameters and necessary information in a sample dataset. This type of description should be included in the dataset documentation.

CASES 2003-2004. ALL nine legs : 08 September 2003 to 26 August 2004.
File created: August 9, 2005 (R2: Release 2)
Last modified: January 11, 2008
Contents of the ASCII files *.int
A Header with the meta-information has been added

Col	Content	Format	Units
1	Pressure (or depth)	F7.2	dbars
2	Temperature	F7.3	deg C (ITS-90)
3	Transmissivity	F7.2	%
4	Fluorescence	F6.2	micrograms / l
5	Salinity	F7.3	PSS 1978
6	Density; sigma(S,T,P)	F7.2	kg/m3
7	Specific volume anomaly	F4.0	10**(-8) m3 / kg
8	N2 : Brunt-Väisälä frequency	E9.2	1/sec2
9	Density; sigma-t: sigma(S,T,0)	F7.3	kg/m3
10	Potential temperature (theta)	F7.3	deg C
11	Sigma-theta: sigma(S,theta,0)	F7.3	kg/m3
12	Freezing temperature	F7.2	deg C
13	Dissolved oxygen concentration	F7.4	ml/l
14	pH	F5.3	no units
15	Nitrates	F8.3	mmol/m3
16	PAR pressure	F7.2	dbars
17	PAR	F8.3	µEinsteins/m2/sec
18	Surface PAR	F8.3	µEinsteins/m2/sec

Data in columns 1, 2, 3, 4, 5, 13, 14, 15 and 16 have been averaged every exact db (pressure). The other parameters have been computed with the averaged data (i.e. pressure, temperature and salinity). The columns show the correct number of significant figures.

NaN indicates Not-a-Number, i.e. no data at that depth (for the first few meters only). At depth, missing values (rare) have been replaced by linearly interpolating the data from the preceding and following depths.

Brunt-Väisälä frequency (N2) has been calculated in a leap-frog fashion: i.e. with the data (potential density) from the previous and following depths; i.e. N2 at 5 m is computed with the data at 4 and 6 m.

Transmissivity, fluorescence, and pH data are the bin averaged, raw values.

Units of the Seapoint Chlorophyll Fluorometer data are micrograms per liter. The sensor has a minimum detectable level of 0.02 and a range of 0-150 micrograms per liter. No calibration check has been performed. Manufacturer coefficients were used.

Dissolved oxygen data has been processed in the following manner. The calibration was first checked by comparing Winkler values with CTD-DO values in the same bottle. The DO sensor calibration coefficients were then corrected. Finally, a delay of x seconds was added to the DO-CTD data to better fit the upcast and the downcast. A different value is used for each leg of the cruise. The data is then averaged in 1 m bins. Only the downcast is given here. The oxygen gradient in regions of high temperature gradient is still difficult to match to the temperature gradient, but this is far from being unusual according to the manufacturer (SeabBird).

PAR pressure is the average pressure (depth) at which the PAR measurements were obtained. The PAR sensor was located 1.71 meters above the CTD pressure sensor.

Missing depths are considered "bad data" by the quality control process. They were removed before the final calculations (density, N2, etc). However, they are available upon request. The quality control process is described in the following document: CASES_postprocessing.doc

Please read accompanying documentation. Also available for each leg is a document titled: Dissolved Oxygen tests to configure Sea-Bird data Processing for Cruise xxxx.pdf

---

[Adapted from Gratton *et al*. for various CASES CTD data. Metadata records available on http://www.polardata.ca and using CASES CTD as keyword]

Appendix C - Example of Alternative Data Arrangement

This arrangement may be more useful when many records do not have measurements for most parameters and thus have many missing values. In this arrangement, one column is used to define the parameter and another column is used for the value of the parameter. Other columns may be used for site, date, treatment, units of measure, etc., as in the example below.

Coast redwood NPP data from Humboldt Redwoods State Park, California, USA; Busing & Fujimori, June 2005

Old stand plot study at Bull Creek with bole diameter measurements at 1.7 m aboveground in 1972 and 2001

Orig_sort_order	Parameter	Measurement_Type	Value	Units	Species	Sequoia_sp_grav	Equation
1	Latitude (N)	Site Characteristics	40.35	decimal degree	NA	-999.9	NA
2	Longitude (E)	Site Characteristics	-123	decimal degree	NA	-999.9	NA
3	Terrain	Site Characteristics	Alluvial flat	Not applicable	NA	-999.9	NA
4	Slope	Site Characteristics	0	degree	NA	-999.9	NA
5	Elevation (above mean sea level)	Site Characteristics	80	m (meter)	NA	-999.9	NA
6	Total site area	Site Characteristics	1.44	ha (hectare)	NA	-999.9	NA
7	Density	Density	380	stems/ha (stems per hectare)	All species	-999.9	NA
8	Basal area	Area	330	m2/ha (square meter per hectare)	All species	-999.9	NA
…	Basal area	Area	329	m2/ha (square meter per hectare)	Sequoia	-999.9	NA
124	Total tree ANPP	ANPP	669-802	g/m2/yr (gram per square meter per year)	All species	0.38	eq. 2 estimates

Sequoia_sp_grav: *Specific gravity, 0.33 mg/cm3, see WE Westman & RH Whittaker, 1975, J. Ecol. for details.

Sequoia_sp_grav: ^Specific gravity, 0.38 mg/cm3, from DW Green et al., 1999, USDA Forest Service FPL-GTR-113.

Method: **Calculations & allometric equations described by RT Busing & T Fujimori, 2005, Plant Ecol.

Notes: ***Range of values results from min. & max. estimation ratios of WE Westman & RH Whittaker, 1975, J. Ecol.

[From: Busing, R. T., and T. Fujimori. 2005. NPP Temperate Forest: Humboldt Redwoods State Park, California, U.S.A., 1972-2001. Dataset. Available on-line at http://www.daac.ornl.gov, from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A.]

Appendix D - README file template

This page may be saved as an TEXT file named "README_{Metadata/Dataset title}_{Today's date}.txt" and submitted along with the dataset files. The outline below should be completed with information relevant to the submitted dataset:

Mandatory information:

1. File names, directory structure (for complex datasets), and brief description of each file or file type:
2. Definitions of acronyms, site abbreviations, or other project-specific designations used in the data file names or documentation files, if applicable:
3. Definitions of special codes, variable classes, GIS coverage attributes, etc. used in the data files themselves, including codes for missing data values, if applicable:
4. Description of the parameters/variables (column headings in the data files) and units of measure for each parameter/variable:
5. Uncertainty, precision, and accuracy of measurements, if known:
6. Environmental conditions, if appropriate (e.g., cloud cover, atmospheric influences, etc.):
7. Method(s) for processing data, if data other than raw data are being contributed:
8. Standards or calibrations that were used:
9. Specialized software (including version number) used to prepare and/or needed to read the dataset, if applicable:
10. Quality assurance and quality control that have been applied, if applicable:
11. Known problems that limit the data's use or other caveats (e.g., uncertainty, sampling problems, blanks, QC samples):
12. Date dataset was last modified:
13. Related or ancillary datasets outside of this dataset, if applicable:

Optional information:

14. Methodology for sample treatment and/or analysis, if applicable:
15. Example records for each data file (or file type):
16. Files names of other documentation that are being submitted along with the data and that would be helpful to a secondary data user, such as pertinent field notes or other companion files, publications, etc.: