BCO-DMO

Data Management Guidelines Manual

a collection of best practice recommendations for collecting and sharing biogeochemical and ecological oceanographic data and metadata

TABLE OF CONTENTS

Executive Summary

The purpose of this document is to provide ocean scientists with a description of "best practices" designed to enable projects to quickly and efficiently make their data publicly available through the auspices of the Biological and Chemical Oceanography Data Management Office (BCO-DMO) (http://www.bco-dmo.org). This office was created to serve Principal Investigators (PIs) funded by the National Science Foundation's (NSF) Biological and Chemical Oceanography Sections as a facility where marine biogeochemical and ecological data and information developed in the course of scientific research can easily be disseminated, protected, and stored for short and intermediate time frames.

The Data Management Office will work with principal investigators to get their data online and make them publicly available. In addition, this office will assist the investigators with data quality control, maintaining an inventory and program thesaurus of carefully defined field names, generating metadata (e.g. Directory Interchange Format (DIF)) records required by Federal agencies, and ensure submission of data to national data centers. The BCO-DMO will support and encourage data synthesis by providing new, online Web-based display tools, facilitate interoperability among different data portals, and facilitate regional, national, and international data and information exchange.

In order to effectively serve data collected in the course of field experiments, it is essential that information about the project objectives, course of events in the field during the data acquisition (Cruise Report and Event Log), description of the equipment and sensors used, and protocols that governed the data acquisition and subsequent data processing be provided. Analogous information is needed for laboratory experimental work or modeling projects.

Data from completed projects should be contributed to the BCO-DMO or served from a project's local machine using the JGOFS/GLOBEC software or compatible software installed on its machine. The data may be contributed in a spreadsheet, in a word processing tabular format, or any computer readable file (except PDF files) in either ASCII or binary format. Supporting documentation (metadata), sufficient to enable others to use the data, should be contributed with the data. See the "Metadata" section of this report for details. Images and video may be contributed in several formats.

For all matters relating to data contribution and accessibility, please contact the Data Management Office at cchandler@whoi.edu or rgroman@whoi.edu, or at http://www.bco-dmo.org/contact.

Introduction

Scientific Investigators funded by the National Science Foundation (NSF) have several reporting obligations with respect to data collected, generated, or analyzed during their funded project. These obligations, in turn, encourage the use of good data management practices and procedures. The following sections specify these obligations, practices and procedures. We recognize that it is nearly impossible to dictate what steps scientific investigators should take during the planning, collection, and processing of their experiments to ensure that their data can be used by others. We do hope, however, that the recommendations and suggestions contained in this guide prove useful. If they are followed, then the storing, retrieval and subsequent reuse of data will be significantly enhanced.

The BCO-DMO (http://www.bco-dmo.org) was created to serve PIs funded by the NSF Biological and Chemical Oceanography Sections as a facility where marine biogeochemical and ecological data and information developed in the course of scientific research can easily be disseminated, protected, and stored for short and intermediate time frames. The Data Management Office also strives to provide research scientists and others with the tools and systems necessary to work with marine biogeochemical and ecological data from heterogeneous sources with increased efficacy. To accomplish this, two data management offices (former- U.S. JGOFS and U.S. GLOBEC) were united and enhanced to provide a venue for submission of electronic data/metadata and other information for open distribution via the World Wide Web. The JGOFS/GLOBEC Client/Server distributed data management system software is used to serve data and information to every investigator, regardless of computing platform. In addition, Web services are provided for data discovery and development is underway of a machine-to-machine application programming interface (API) that will allow interoperability between data systems on the Web. The BCO-DMO will manage existing and new data sets from individual scientific investigators, collaborative groups of investigators, and data management offices of larger multi-institutional projects via any standard Web browser. This office will work with principal investigators on data quality control, maintain an inventory and program thesaurus of carefully defined field names, generate metadata (e.g. Directory Interchange Format (DIF)) records required by Federal agencies, and ensure submission of data to national data centers. The BCO-DMO will support and encourage data synthesis by providing new, online Web-based display tools, facilitate interoperability among different data portals, and facilitate regional, national, and international data and information exchange.

It is important to emphasize from the beginning that the intellectual investment and time committed to the collection of a data set entitles the investigator to the fundamental benefits of the data set. Publication of descriptive or interpretive results derived immediately and directly from the data is the privilege and responsibility of the investigators who collect the data within the time period set by the investigator's program manager. Any person making substantial use of a data set must communicate with the investigators who acquired the data prior to publication and anticipate that the data collectors will be co-authors of published results. This extends to model results and to data organized for retrospective studies.

Data Acquisition Procedures and Protocols

The fundamental objectives of the BCO-DMO are dependent upon the cooperation of scientists from all the participating disciplines. Physicists, biologists, and chemists must make use of data collected during their field programs, laboratory experiments, as well as historical studies to further our scientific understanding. The scientific objectives require quantitative analysis of interdisciplinary data sets and therefore data must be exchanged among researchers. To extract the full scientific value, data must be made available to the scientific community on a timely basis and be accompanied by sufficient documentation to support accurate reuse.

All principal investigators, via their proposals, submit plans for the collection and processing of data. A single description of the expected data sets, a "data plan", will be derived by the Data Management Office from the submitted proposals. Where a group of investigators are cooperating in managing and collected data, a single responsible scientist should be identified for each measurement.

To provide the opportunity for comparison with historical data, measurement techniques should be consistent with techniques used to collect the existing data unless there is significant scientific justification for change. When new techniques are adopted, methods/protocols for relating the new data to existing data should be developed.

In order to allow for comparison of different data sets, the principal investigators should document the procedures/protocols that will be used to collect and process samples and data. For large programs, a data policy describing sampling and analytic protocols should be published and shared among participating investigators. For additional information about the U.S. GLOBEC Data Policy, see *U.S. GLOBEC Data Policy, U.S. Global Ocean Ecosystems Dynamics Report Number 10, February 1994* from which much of the above material is extracted. The US JGOFS Data Policy is published online, http://usjgofs.whoi.edu/jgofs-data-policy.html, and the US JGOFS protocols were originally published as Protocols for the Joint Global Ocean Flux Study (JGOFS) Core Measurements, 1994 UNESCO, and subsequently as JGOFS Report No. 19. (1996) and are available online (http://usjgofs.whoi.edu/protocols.html

Metadata and Data

Data sets consist of both the actual measurements and also descriptive data, generally referred to as metadata. The distinction between data and metadata is blurred; some say “it’s all data”. However, the generally accepted view is that metadata comprise location, time, units, accuracy, precision, methods of measurement and sampling (sampling and analytical methodologies), investigator, reference to publications describing the data set, and a description/protocol of the processing of the data collected. Other metadata related to cruises include

- Name of the ship
- Investigator-designed cruise identifier (e.g. “FN9801 Leg 1”)
- Associated project
- Associated Institute
- Principal investigator(s) for the cruise and their contact information
- Other responsible investigators, and their contact information

Principal investigators are responsible for the selection of methods, equipment, and calibration procedures. If new measurement techniques are used, methods relating the new data to existing data should be developed and documented. Whether your measurement and processing techniques are new or old, they need to be provided to the DMO so that this information can be made available to others using the collected data.

There are several approaches to recording the necessary metadata for stations, net tows, and bottle casts. Sometimes such information is recorded in the Cruise Reports; other times, this information may be part of the digital data supplied with the sample results. Appendix 4 gives examples of the types of information that should be preserved and documented.

Metadata for a particular data set describe, for example, how, when, where, and by whom the data were collected, created, accessed, and/or modified and how they are formatted. Metadata describe the content, quality, condition, and other characteristics of data. Metadata are essential for understanding the data set and are vital in helping potential users find needed data and determine whether a data set will meet their needs.

In addition to metadata that describe each data set, multi-disciplinary, oceanographic research programs benefit from metadata that describe each sampling platform. Most commonly this will be a research cruise, but might also include moorings, sediment traps or mesocosm experiments.

Cruise level metadata components are the responsibility of the Chief Scientist and include:

1. cruise plan (pre-cruise document describing planned activities)
2. cruise report (document describing work done during cruise)
3. cruise summary descriptive fields and data inventory list
4. event log (a sequential list of all measurement activities during the cruise)

Cruises

By “Cruises”, of course, we are merely referring to all sorts of human endeavors that take place on sea-going sampling platforms.

Cruise Plan

The cruise plan is a document summarizing the expected operations during a planned cruise. The WOCE program encouraged chief scientists to submit a cruise plan one year prior to the cruise, and found it to be a useful document. For coordinated programs (e.g. GEOTRACES) the usefulness of such a document makes them essential. Many ship operators require a cruise synopsis which often includes much of the information suggested for inclusion in a cruise plan document.

Information to be specified in the cruise plan includes:

- Cruise name (unique designator, e.g. ship code and number)
- Ship name
- Chief Scientist name
- Dates
- Ports of call
- Description of scientific goals, including proposed cruise track and sampling plan
- Expectations for data to be collected (identified by responsible PI, other personnel and their respective affiliations)
- List of standard shipboard (underway) measurements
- List of sampling instrumentation

Report

A project report, describing what was accomplished during the project, is typically required by NSF in the form of annual reports and a final report. We have found that when research projects involve a cruise, Cruise Reports generated during the cruise and finished very soon after the cruise is completed, substantially facilitate subsequent use of the data collected during the cruise. These reports describe in more detail than an annual or final report what happened during a cruise, what data were collected, what problems were encountered during the data collection phase of the project, and how they were overcome. We strongly recommend that the Chief Scientists coordinate the preparation and completion of cruise reports and send a digital copy of the report to the Data Management Office (DMO). The BCO-DMO will incorporate the cruise report on the web site as integral information about the data collected. A sample cruise report Table of Contents (report outline) page is shown in Appendix 1.2.

Summary

If a full cruise plan document is not created, then it is useful to at least document the basic cruise metadata as a list of all expected results from the sampling program (often referred to as the data inventory). In addition to the initial fields listed below, the cruise summary includes a list of all PI and co-PI names, their contact information, and a brief description of all expected measurements and associated instrumentation. Each individual’s affiliation, *i.e.* institution or organization, at the time of the cruise is also a key piece of information for the cruise summary.

Event Log

The cruise event log is a data set and many consider it to be an essential part of the cruise metadata (data about the cruise). Many investigators create and maintain the event log as an Excel spreadsheet. It is important that the event log be as complete and accurate as possible. The event numbers will allow investigators to integrate data from different sampling devices used during a cruise, or combine data from like sampling devices from multiple cruises. For inter-disciplinary projects gathering data from multiple sampling devices, the event log can dramatically improve the ability to integrate discrete data sets contributed by participating investigators. See Table 1 for a suggested list of fields. We encourage every Chief Scientist to generate a cruise event log that records all the sampling events during a cruise.

The event log provides an overall summary of the activities during a cruise. If your project involves going out to sea, we strongly recommend that an event log be generated during the cruise. The BCO-DMO can provide an Excel spreadsheet program (events.xls) that can be used to help prepare the event log in the suggested format. However, you are free to generate the log in any way you choose. The BCO-DMO event log includes the field names described in Table 1, although new ones can be added and incompatible ones deleted.

While maintaining the event log it is important to note the following: Breaking off and resumption of operations on cruises should be logged as events in the event log, using the s/e (i.e. start/end flag) column. Also, aborted tows (that is, tows that result in no data) should **not** be logged. Rather, indicate in the comment field of the entry for the successful tow that the previous attempt was aborted.

The event log should be included in the cruise report and made available online. The event log should be updated daily by cruise participants while at sea and carefully verified. Additionally, a technique that works well is to appoint a single person whose responsibility it is to check the event log every 12 hours and review entries from the previous 24 hours. This technique affords a timely opportunity to correct any missing or questionable entries. The final version of the event log should be sent to the DMO in Excel, comma separated, or "flat file" ASCII format immediately after the end of the cruise for posting online. A sample event log is shown in Appendix 1.4.

As a reminder, each ship collects so-called along track data, e.g. date, time, latitude, longitude, and data from each of the ship's sensors such as wind velocity, water depth, air temperature, sea surface temperature, water currents (ADCP), etc. Make sure to get a copy of these data and send a copy to the Data Management Office immediately after your cruise.

Table 1: *A list of suggested field names and their descriptions is shown for a typical cruise event log. We highly recommend that large programs develop a customized spreadsheet (e.g. Excel) that can be used to generate consistent event logs for all cruises. In any case, investigators are encouraged to design event log schemas that will support their research objectives in addition to providing a record of the necessary sampling event data. Event logs should reflect program-specific sampling schemas or considerations (e.g. standard station or sampling grid identifiers, relative position with respect to environmental features such as eddy location, river outflow, or patch of experimentally treated water).*

Fields in bold are required. Additionally, the event log must also include time in either GMT or local time, and either a station or a cast number if appropriate to the sampling event.

Field Name	Description
event	The unique event number entry can be created by you or perhaps by an application aboard ship. The objective is to create a unique identifier for each sampling event (start and end). One example: shipYYddd.### where ship = ship abbreviation (usually up to 4 characters) YY = last 2 digits of year ddd = day of the year (001 is January 1) ### = consecutive number within the day, starting at 001 each new day. or shipYYMMDD.hhmm where MM is a 2 digit month, DD the 2 digit day, hh the 2 digit hour and mm the 2 digit minutes
instrument	Instrument name (e.g. CTD, TM, MOC10)
station	Station number (assigned by you)
cast	Cast number (assigned by you)
ev_type	Event type descriptor (e.g. a combination of instrument and cast number)
date (GMT)	Date as YearMonthDay the measurement was taken (8 digits as YYYYMMDD) Year as YYYY, Month as MM (01-12) and Day as DD (01-31); January 1, 2008 would be 20080101
time (GMT)	Time (24 hour clock, as hhmm with hh begin the two digit hour and mm being the two digit minute. Fractions of a minute can be represented as mm.mmmm, if necessary.
time_local	(optional) local time as hhmm or hhmm.mm (similar to time format)
time_zone	Time zone using any standard convention as long as you describe what convention you are using. For example, for California Coast, the time zone can be entered as UTC-08, +8, or PST.
s_e	Start/end flag of the activity. Enter s for the start and e for the end. This might also include a flag (m) for maximum depth in the case of vertical profiles.
latitude	Latitude. Positive values mean the northern hemisphere and negative values mean the southern hemisphere. Preferred format is sdd.dddd for decimal degrees where s is a minus sign (plus sign is optional).
longitude	Longitude. Positive values mean east of the prime

	meridian and negative values mean West of the prime meridian. Preferred format is sddd.dddd for decimal degrees where s is a minus sign (plus sign is optional).
water_depth	Water depth in meters
cast_depth	Cast depth in meters
investigator	Name of person responsible for sampling event
custom program-specific fields	e.g. standard station or sampling grid identifier, distance to feature
comments	Optional comments

Moorings, Gliders, and Laboratories

To Be Added (TBA)

Modeling

TBA

Publishing the Data

Serving the Data

Because the program uses the JGOFS/GLOBEC distributed data management software to provide access to data, your data can be made available to others by either "serving" the data yourself or by sending them to the DMO for serving by the BCO-DMO data server. Contact the BCO-DMO staff when you are ready to contribute your data for publication on the server.

Long Term Data Archiving

The Data Management Office can initiate or facilitate the transfer of your data to the appropriate national data center for permanent archiving.

Appendices

Appendix 1: Examples for Cruises

1.1 Sample Cruise Plan

1.2 Sample Cruise Report (just the Table of Contents)

R/V ALBATROSS IV AL9906 Cruise Report

TABLE OF CONTENTS

Full Report: http://globec.whoi.edu/globec-dir/reports/al9906/al9906rpt.html

1.3 Sample Cruise Summary

An example of cruise summary metadata:

```
chief scientist: Dennis McGillicuddy
affiliation:     Woods Hole Oceanographic Institution (WHOI)
funding:         NSF: OCE0241310
cruise dates:    11 June 2004 to 03 July 2004
ports:           Woods Hole, MA to St. Georges, Bermuda
program:         EDDIES
project:         2004 Survey 1
platform:        R/V Oceanus
cruise:          OC404-1
location:        Sargasso Sea
related:         ops coordinated with R/V WEATHERBIRD II
```

Example of the expected data inventory contributed with the cruise summary:

version 21 November 2007
EDDIES Cruise and Data Inventory
#
Expected EDDIES data contributions
R/V OCEANUS 2004 Survey 1 cruise: OC404-1
#

Instrument	Measurement	PI_name	PI_Affiliation
Athena	Ship track (XYT data)	McGillicuddy	WHOI
manual	Cruise report	McGillicuddy	WHOI
manual	orig sample ID logs (scanned, PDF)	McGillicuddy	WHOI
manual	cruise event log	McGillicuddy	WHOI
SeaBird CTD	CTD profiles	McGillicuddy	WHOI
Niskin	bottle file (base)	McGillicuddy	WHOI
Niskin	bottle salinity	McGillicuddy	WHOI
Niskin	Nutrients NO3/PO4/SiO2	McGillicuddy	WHOI
Niskin	bottle oxygen	Bates	BIOS
Niskin	nM NO3/PO4	Hansell	RSMAS
TM Rosette	3He; Tritium	Jenkins	WHOI
Niskin	Chlorophyll	Bates	BIOS
Niskin	HPLC pigments	McGillicuddy	WHOI
Niskin	DIC	Bates	BIOS
Niskin	pCO2	Bates	BIOS

ADCP	underway ADCP	McGillicuddy	WHOI
Athena	underway Athena	McGillicuddy	WHOI

1.4 Sample Event Log

This sample event log is an example of entries for some common event log fields. The latitude and longitude entries follow the convention of indicating hemisphere as positive for North and East and negative for South and West. The 'instrument' entries are most useful if the instrument names are selected from a list of names for sampling instrument types deployed during the cruise.

event	date	GMT	time_L	TZ	station	longitude	latitude	instrument	investigator	S_E	ev_type
MV_0220400	20020122	0400	1700	13	TEST	-55.400	-171.470	CTD	nd	S	CTD001
MV_0220430	20020122	0430	1730	13	TEST	-55.402	-171.472	CTD	nd	E	CTD001
MV_0230442	20020123	0442	1742	13	1	-55.398	-171.480	CTD	Wang	S	CTD002
MV_0230642	20020123	0642	1942	13	1	-55.396	-171.481	CTD	Wang	E	CTD002
MV_0231556	20020123	1556	0456	13	1	-55.407	-171.583	ZooTow	Landry	S	Zooplankton Tow
MV_0231800	20020123	1800	0700	13	1	-55.405	-171.581	ZooTow	Landry	E	Zooplankton Tow
MV_0240153	20020124	0153	1453	13	1	-55.329	-171.490	TM	Wang	S	TM001
MV_0240153	20020124	0420	1720	13	1	-55.330	-171.492	TM	Wang	E	TM001
MV_0240745	20020124	0745	2045	13	1	-55.335	-171.408	ThPump	Andrews	S	Pump Cast 01
MV_0240745	20020124	0920	2220	13	1	-55.334	-171.409	ThPump	Andrews	E	Pump Cast 01

Appendix 2: Examples for Moorings, etc.

Appendix 3: Examples for Modeling Results

Appendix 4: Metadata

4.1 Suggested metadata fields for a data set

The fields below represent many of the most common metadata fields important to preserving high quality marine chemistry data. These fields are not necessarily complete and investigators are encouraged to include additional information required to describe their data. The intent is to describe the general types of information that should be contributed as supporting documentation with the data. Documentation of quality control procedures is especially valuable. Also, note the final line below. It is essential that the missing or 'no data' indicator be explicitly stated. For example, it must be clear whether a data value of zero indicates a quantity of zero for that parameter as opposed to just 'not measured'. Likewise, empty data fields (nothing entered) should be avoided as this just leads to ambiguity.

This is an example of metadata that describe a trace element data set.

```
Data set name: trace Fe and Al

 PI name:         Christopher Measures
 affiliation:     University of Hawaii
 dataset:         Dissolved iron and aluminum concentrations from TM casts
 project/cruise:  Arabian Sea/TTN-043 - Process Cruise 1
 ship:            R/V Thomas Thompson
```

PI Notes and Methodology [include a separate document describing sampling and analytical methodology, quality control measures and any comments from PI that support these data]

Parameter	Description	Units
event	sampling event number from event log	dimensionless
sta	station number from event log	dimensionless
sta_std	Arabian Sea standard station identifier	dimensionless
cast	cast number from event log	dimensionless
bot	Trace Metal-free Go-Flo bottle number	dimensionless
depth	sample depth observed	meters
AlFe_filt_sz	non-zero indicates filter size	microns
Fe_unfilt	iron concentration, unfiltered	nanomoles/liter
Al_unfilt	aluminum concentration, unfiltered	nanomoles/liter
Fe_diss_lt0d2	dissolved iron conc. <0.2 microns	nanomoles/liter
Al_diss_lt0d2	dissolved aluminum conc. <0.2 microns	nanomoles/liter

nd indicates missing or no data

4.2 Metadata fields for Station, Net Tows, and Bottle Casts

The fields below represent many of the most common metadata fields important to preserving high quality plankton data. These fields are not necessarily complete nor are all fields mandatory. Many of these data fields may be better stored along with each station or tow within the plankton data sheet itself. The intent is to describe the general types of information that should be preserved either within the data or as a separate metadata or cruise summary.

[This example is from the Appendix 3 in the ICES Working Group on Zooplankton Ecology (WGZE) report, 24 February 2003 (ICES CM 2003/C:01).

Metadata-fields relating to a specific STATION:

This is specific information relating to the position and time of the sampling station, along with the weather conditions and other details observed during the sampling.

Station latitude and longitude (noting any hemisphere indicators such a “N” for North or negative (-) for South, etc.)
Station Date (Month, Day, Year)
Station Time (designated as “local”, “GMT/UTC”, “ship”, etc)
Investigator-designated **Station Identifier** (e.g., “Station 1x”, “Station 2x”, ...)
Optional general station time (twilight, midnight, day, morning)
Meteorological Observations (windy, wavy, cloudy, sunny)
Station Sounding (bottom depth)

Information about any other supplementary/complementary data collected at the same time (same station) should also be supplied (i.e., a note that “CTD and nutrient samples were also made at this station”)

Note any affecting instances or corrections applied (e.g., “a substantial phytoplankton bloom was present at this station” or “a larger net mesh size was used at this station due to frequent clogging by gelatinous zooplankton”)

Metadata-fields relating to the NET TOW or BOTTLE CAST:

This information describes the towing (or bottle deployment) methods and procedures.

Towing Method (horizontal, vertical, oblique)
Towing depth-range (a range of starting and ending depths for each net or bottle), or the wire angle and wire out during the tow
Towing Duration (minutes or hours)
Towing Distance (in meters)
Average Towing Speed (knots or meters per second)
Note any affecting instances or corrections applied (e.g., “the gear hit the bottom midway through the tow”)

Metadata-fields relating to the SAMPLING GEAR:

This information describes the sampling gear employed, with key metadata fields such as the effective mesh size of the sampler.

Describe the sampling gear used, providing a literature reference if available
If using a “standard” net (e.g., a NORPAC net) was used, be sure to note any modifications to this net
What net mesh size was used (usually in microns)
What was the net opening shape (square or circular) and the opening mouth area or diameter
Was a flowmeter used? When and how was it calibrated?
Note any affecting instances or corrections applied (e.g., did the flowmeter break or the codend crack)

Metadata-fields relating to SAMPLE PROCESSING:

This information describes the sample processing methods and protocols.

What volume of water was filtered to yield this sample (i.e., from the flowmeter or calculated via mouth area and towing distance, or estimated because the flowmeter failed)?

How were samples preserved, and in what (e.g., 5% buffered formalin)?

How were samples processed (summarize the counting, weight, or volume method)?

Was the sampled split (via Folsom splitter or other method)? What was the size of the final aliquot?

Were large plankters removed prior to making biomass measurements? Was a size or volume criteria used in deciding what to remove and what could remain?

Investigator-designated tow, net, or sample identifier (e.g., "Sample 1035 from Net 5")

Note any affecting instances or corrections applied (e.g. "eggs and fragments were not counted")

Metadata-fields relating to SAMPLE ITSELF:

This measured plankton data fields should be clear enough to be understood by others with slightly less expertise and situational knowledge than the original investigator. The two most common mistakes are not providing units for each measurement (e.g., "number per cubic meter" or "milligrams wet mass per sample"), and not providing clear column headings for the data (e.g., what is "CfcV" and "HetBact"?)

Provide the units for each measurement (e.g., #/liter, #/m3, #/m2, mg/m3, mg/haul, ...)

If taxonomic codes, symbols, or abbreviations are used in the data, provide a translation table to help reduce possible misunderstandings of the taxa (e.g., "CfcV" = "*Calanus finmarchius* copepodite V", ...)

Is an estimate of final uncertainty of the data known?