Guidance on providing data to BODC for Shelf Sea Biogeochemistry (SSB): version 2

Updated: 27 May 2014
Updated to include links to the SSB website

This document provides guidance on submitting data to the British Oceanographic Data Centre (BODC) for Shelf Sea Biogeochemistry (SSB) project participants. The guidelines conform to good practice criteria adopted across the NERC Environmental Data Centres. For more information, please contact the SSB data managers, Louise Darroch (lorr@bodc.ac.uk) and Sean Gaffney (sgaf@bodc.ac.uk).

1 Contents

2 Supporting data set documentation

All data sets submitted to BODC must be accompanied by supporting documentation. A BODC data set documentation template to accompany environmental data submissions is provided in Appendix A.

3 Guidance Notes for providing data

The key thing to aim for is that the data set and documentation should allow the data set to stand on its own and be useable without returning to the data set author(s).

3.1 Formats

The format must be well-documented and conformant with widely accepted standards, such as ASCII (which includes .csv) or NetCDF. Preferably data formats should conform to internationally-agreed content standards, such as CF-compliant NetCDF or SeaDataNet ASCII spreadsheet format. The format must be readable by tools that are freely available now and are likely to remain freely available indefinitely. Formats can be 1-D, 2-D, 3-D etc.

3.2 Data file header contents

In order for all information to be captured and explained, so that data users should not necessarily need to return to the data author with simple queries, the following information should be included with the data. A simple way to do this is for it to be incorporated into a header in the data file (see example in Appendix B). The fields suggested below are a suggested minimum, if there are further information relevant to your data set that you feel are not covered then please add them to the header or the associated documentation. For example, this might include alphanumeric codes used to identify a station, physical sample or experiment (see 'Labelling and metadata standards for Shelf Sea Biogeochemistry' for more information on alphanumeric codes). Descriptions are given for each entry suggested below:

Author(s):	To which the dataset will be attributed
Institute(s):	e.g. PML, MBA, SAHFOS
Contact email:	
Data collected by:	List of people involved in collection, analysis, production of the data set beyond those listed as authors.
Project PI:	
Project:	Project and work package which the data belong to

Funding Body:	
Dataset title:	
Parameter Names (if shortened versions used in column titles):	Description of any short hand data column titles e.g. Temp = Temperature of the water column measured by a CTD; Sal = Salinity of the water column measured by a CTD; … etc. If there are a large number of columns a separate table in the documentation may be more user friendly (see below for an example).
Parameter Units:	Description of any short hand units e.g. Temp = degrees Celsius (ITS-90); Sal = dimensionless; … etc. If there are a large number of columns a separate table in the documentation may be more user friendly (see below for an example).
Description of parameter flags in column QF (if used):	If established data flags are used quote what source e.g. WOCE (http://woce.nodc.noaa.gov/woce_v3/wocedata_1/woce-uot/document/qcflags.htm), BODC (http://www.bodc.ac.uk/data/codes_and_formats/request_format/), etc… . If using your own system provide a description here (e.g. < = less than detection limit; M = suspect data; T = interpolated data etc).
Absent data value:	Rather than leaving cells blank, which can leave doubt over whether this was intentional or an oversight, it is sensible for an absent data value to be used. Additionally a flag (e.g. "N") can be used in connection with an empty cell. Please take care when selecting an absent data value, it should be a value that is not likely for a parameter (e.g. for SST -1 is a poor choice, whereas -999 is a good example).
Dataset version:	If submitting for a DOI should be "FINAL". Nothing to stop use of this template for your own day to day data management and this field can help you keep track of the version e.g. "Preliminary", "Pre-QC" as you chose.
Submission date:	Date sent to BODC
Associated descriptive document filename:	Name of associated document containing abstract, sampling protocols and overall comment on data quality.

3.3 Metadata columns

For illustration, the metadata columns are highlighted in green and blue in the example shown in Appendix B. Below is the suggested minimum there should be to accompany marine series and sample data:

- Date/time (in GMT/UT) - if no accurate time is available add an adjacent column and flag 'time approx'
- Latitude (decimalized +ve = North) - if no accurate lat is available add an adjacent column and flag 'lat approx',
- Longitude (decimalized +ve = East) - if no accurate lon is available add an adjacent column and flag 'lon approx',
- Sample or measurement depth (m),
- Vessel the measurements or samples were collected on (or cruise ID),
- Sampling gear or instrumentation used (e.g. VMADCP, CTD rosette, Niskin, Bucket, Surface Pump, Nets, Grab, Corer, etc),
- Fixed station name (site name), if a recognised site is being sampled.

For Shelf Sea Biogeochemistry, files should include metadata columns for the alphanumeric codes to identify a station (and cast), physical sample (or sub-division) or fixed station as shown in Appendix B (highlighted in green) (see the document, '[Labelling and metadata standards for Shelf Sea Biogeochemistry]' for more information on alphanumeric codes). Where this information is the same for all the data represented in the data set it only needs to be included in the header once.

Most series data will only require the suggested minimum above. For sample data, the metadata columns may need to include more options depending on the sampling gear used to collect the data (see below for additional fields for each sample gear). Where this information is the same for all the data represented in the data set it only needs to be included once in the documentation (e.g. "The data set is from the analysis of a repeated series of vertical hauls at 1 m/s from 50 m to the surface using a 200 um mesh bongo net, 0.5m diameter, opening area ~0.196 m^2, filtering ~9.82 m^3 of water on each haul") or once in the header of the file.

Bucket, Pump, Niskin, Corer, Grab, Marine Snow Catcher	Minimum above		

CTD rosette	Minimum above plus		
	Rosette bottle	if available	
	Firing Sequence	if available	

Nets/trawl	Minimum above plus		
	Mesh size		
	Opening area		
	Haul type	e.g. Vertical, horizontal, oblique	
		For a vertical haul	Minimum depth
			Maximum depth

		For a horizontal trawl	Start time, lat & lon
			End time, lat & lon
			Depth, 'surface' or 'benthic'
	Haul/trawl speed		
	Volume filtered		
Sediment traps	Minimum above plus		
	Sample cup ID	For each cup	Date & time cup opened
			Date & time cup closed

Where field or laboratory incubations or experiments have been carried out, the reference identifying the activity and the details of the sample collection that provided the basis for the activity should be provided in addition to the following metadata columns (see below). For Shelf Sea Biogeochemistry, files should include the alphanumeric codes to identify the incubation or experiment and the physical samples (or sub-divisions) used (see 'Labelling and metadata standards for Shelf Sea Biogeochemistry' for more information on alphanumeric codes). Where metadata or information is the same for all the data represented in the data set it only needs to be included once in the documentation or once in the header of the file.

Incubations	Start & end time	
	Treatment	Light, temp conditions and additions (tracers, nutrients, etc)
	Incubation Volume	
	Time Course Point	Where a series of samples are taken during the incubation at different time steps
Settling velocity tubes	Start & end time	
	Sub-sample Volume	Volume of the sub-sample withdrawn from the vessel
	Time Course Point	Where a series of samples are taken during the resting period at different time steps

3.4 Data columns

These have been highlighted in orange for illustration in Appendix B. Add data columns as required.

3.5 Notes on taxonomic data

Since species nomenclature does change with time it is useful to include World Register of Marine Species IDs (WoRMS AphiaID) (http://marinespecies.org/) and/or Integrated Taxonomic Identification System IDs (ITIS TSN) (http://www.itis.gov/), e.g. *Arenicola* spp. AphiaID = 129206; ITIS TSN = 67507, along with the taxonomic name. This can either be included as additional columns in the parameter description table in the documentation or as additional rows in the data file:

				AphiaID	129206
				IT IS TSN	67507

Lat	Lon	Date and Time	Depth	Taxa	*Arenicola* spp.

3.6 Notes on chemical data

Some chemicals, particularly hyrdocarbon chemicals, can have a number of common names. In such situations the inclusion of the Chemical Abstracts Service Registry Numbers (CASRN) (http://www.cas.org/expertise/cascontent/registry/regsys.html) can be useful in removing ambiguity or confusion. This can either be included as an additional column in the parameter description table in the documentation or as additional rows in the data file:

				CASRN	7314-30-9
Lat	Lon	Date and Time	Depth	Chemical	DMSP

3.7 Notes on field or laboratory incubations or experiments

The data section of these files should contain the time-course of measurements or the initial/final series of measurements in an incubation or experiment. Time points must express the time elapsed since the incubation or experiment began. If the incubation or experiment was used to derive *in-situ* rates, these should also be reported.

3.8 Example of parameter and unit description table

If there are a large number of columns in the data file a separate table in the documentation may be more user friendly (see below for an example).

Column title	ITIS ID	Aphia ID	CASRN	Units	Description
Temp				Degrees Celsius (ITS-90)	Temperature of the water column measured by a CTD
Sal				dimensionless	Salinity of the water column measured by a CTD
Density				kilograms per cubic metre	Density of the water column measured by a CTD
Xmiss				Percent	Transmissance of the water column measured by a transmissometer
PAR				Watts per square metre	Down-welling PAR irradiance in the water column measured by a PAR sensor
***Arenicola* spp.**	67507	129206		abundance per square metre	Abundance of *Arenicola* spp.
DMSP			7314-30-9	nanograms per cubic metre	Concentration of Dimethylsulfoniopropionate (DMSP) in the water column

3.9 File names

Make sure file names are descriptive (e.g. WCO_2011_nutrient_dataset.csv) not generic (e.g. Data_for_BODC.csv). For Shelf Sea Biogeochemistry, it is preferable to use filenames based on the alphanumeric codes to identify a station, physical sample or experiment (e.g. JC090_040_02.csv). Use matching filenames for the documentation and data file, unless the documentation refers to many data files (e.g. in the case of CTD or optics profiles).

3.10 Other comments

When submitting Excel files please save the files as a comma separated values file (.csv) or tab delimited text file for submission. However take care as coloured text, filled cells and comments boxes are lost when saving from Excel to a csv file. For this reason these formatting methods should never be used for flagging data in a final data set submission. Instead each data column should have a data flag column adjacent to it, with flags applied in this column as text rather than formatting. Additionally data columns should only contain numbers not a mixture of text and numbers (see example below).

Nutrient_x (umol/l)
0.04
0.85
0.11
<0.02

should be represented as

Nutrient_x (umol/l)	QF
0.04	M
0.85	M
0.11	
-999	N
0.02	<

For data that are below the detection limit of the technique please include the detection limit in the data column with an appropriate flag in the flag column. Also confirm the detection limit in the documentation.

Please select the most appropriate level of significant figures for the data you are submitting. Be aware that the formatting of numbers in an Excel spreadsheet will be taken on in the csv file as the absolute number (e.g. 2.456789 formatted to 2dp in an Excel worksheet will only be retained as 2.46 in the csv and if you want a higher level of accuracy reported you should set that appropriately before saving as a csv).

Further tips and additional details can be found with the Best Practices for Preparing Environmental Data Sets to Share and Archive (https://daac.ornl.gov/cgi-bin/MDE/S2K/bestprac.html).

4 Appendix A

4.1 BODC data set documentation template to accompany environmental data submissions

Data set title:

Data set creator(s) and institute(s):

Data set period:

yyyy-mm-dd to yyyy-mm-dd

Data set abstract:

Brief explanation of the data set covering the basics who, what, why, when, where and how (more of a general overview of the how in the abstract with more details of the methods included below). This should relate to the dataset not the project.

Sampling methodology and description of analytical techniques:

Further details of the 'how' given in the abstract. Sampling protocols and analytical techniques used with references if appropriate.

Data quality comments:

An over view of the data set quality. Can simply be a sentence "The author does not have any concerns over the quality of the data set and values are consistent with similar data reported in the wider literature" to something more in depth. Whatever you feel is relevant to the data in the submission.

Name of data file(s):

This could be a zip folder containing files in the case of 1 data series per file e.g. CTD or optics profiles. In the case of multiple files please include rational behind individual file naming.

5 Appendix B

5.1 Example of a submission file (CTD discrete data in csv format)

					Author(s):								
					Institute(s):								
					Contact email:								
					Data collected by:								
					Project PI:								
					Project:								
					Funding Body:								
					Dataset title:								
	Parameter Names (if shortened versions used in column titles):												
					Parameter Units:								
					Description of parameter flags in column QF (if used):								
					Absent data value:								
					Dataset version:								
					Submission date:								
					Associated descriptive document filename:								
CRUISECODE	STNNBR	CASTNO	SAMPNO	SITE	Latitude[deg+veN]	Longitude[deg+veE]	yyyy-mm-ddThh24:mi[GMT/UT]	Rosette bottle	Sample depth [m]	Temp [degC]	QF	Sal [dimensionless]	QF