

NOAA/NODC/Ocean Climate Laboratory 'bioxls' Data Digitization Format Structure

1. INTRODUCTION

As part of the Intergovernmental Oceanographic Commission (IOC) Global Oceanographic Data Archaeology and Rescue (GODAR) and World Ocean Database (WOD) projects, the Ocean Climate Laboratory (OCL) has developed a procedure to convert data from hardcopy reports into digital form so that these data can be incorporated in the World Ocean Database more rapidly and efficiently. The data are entered into an Excel spreadsheet, using a flexible format designed at the OCL, and the output is a "comma separated value" (.csv) file. There are two versions of the format. Version 1 was developed in July 1996; version 2 was developed in July 2001.

2. THE FORMAT

Each .csv file contains information for one cruise and consists of four sections:

- (1) CRUISEINFO: contains information to identify an individual profile, such as the NODC country code, the project associated with the data, a cruise number, the name of the platform (ship) from which measurements were made, a cruise number, institution, and principal investigator.
- (2) STATION: contains information about the station at which measurements were taken such as the latitude (in degrees, minutes, and seconds), latitude hemisphere (north or south), longitude (degrees, minutes, and seconds), longitude hemisphere (east or west), the date (month, day, and year), and originator's station number.
- (3) HEADERS: contains all metadata information such as the time, meteorological data, methods description, gear, and bottom depth. If a dataset replaces a previously processed cruise, with the addition of new parameters, then HEADERS section might have NODC accession number, ship, and a unique OCL-assigned station number which identifies each station.
- (4) DETAILS: contains information about the variables, the units, and decimal places.

3. DETAILS FOR EACH OF THE SECTIONS

- (1) CRUISEINFO section:

While the STATION, HEADERS, and DETAILS sections are repeated for every station, the CRUISEINFO section appears only once and is present at the beginning of each data file.

Here is an example of a CRUISEINFO section in the Version 1 format:

```
CRUISEINFO
49    COUNTRY CODE    Japan
656   INSTITUTE       Tokyo University O.R.I (institute that collected data)
      PLATFORM        Hakuho Maru
      CRUISE           KH-78-3
343   PROJECT          JARE (JAPANESE ANTARCTIC RESEARCH EXPEDITION)
3     REFERENCE        Reversing thermometer
289   COUNTING         SOSC (institution where plankton were counted)
289   VOUCHER          SOSC (institution where plankton samples are located)
2     NUMBER OF PI    number of principal investigators
14    943              DONALD S. DAY
-5006 303             RODNEY ADAMS
```

The first row is the label CRUISEINFO. In Version 1, integer codes are entered into column 1 (see below for more information on codes); the labels are in the second column; and text information, if available, would be entered into column 3. In Version 2, the labels are in the first column, the codes are in column 2, and textual descriptions are in column 3.

In Version 1, the NUMBER OF PI row identifies the number of principal investigators (PIs) for the cruise. This is then followed by the appropriate number of rows to identify each PI, in this case, two. The columns containing the PI information include a parameter code in column 1 (if there is not a parameter code, this would be blank); the PI code in column 2; and the PI name in column 3. If the PI code is not available, column 2 will be blank. There should be as many PI entries as the NUMBER OF PI row indicates. In Version 2, the label, NUMBER of PI, has been eliminated and only the PI names and codes have been entered.

Here is an example of a CRUISEINFO section in the Version 2 format:

```
CRUISEINFO
COUNTRY    49      Japan
PLATFORM   4725    HAKUHO-MARU
INSTITUTE  656      TOKYO UNIVERSITY
CRUISE     KH-78-3
PROJECT    343      JARE (JAPANESE ANTARCTIC RESEARCH EXPEDITION)
PI         943      DONALD S. DAY
PI         303      RODNEY ADAMS
```

All codes are available at http://www.nodc.noaa.gov/OC5/WOD/wod_codes.html.

COUNTRY CODE can be found in the “country.txt” file. It is the country that collected the data. A PROJECT code, if there is one, can be found in the “projects.txt” file. INSTITUTE, COUNTING, and VOUCHER can be found in the “inst.txt” file. The INSTITUTE refers to the

institute that collected the data. The COUNTING code, if there is one (0 or left blank if there is no code), refers to the institution which counted plankton. The VOUCHER code, if there is one (0 or left blank if there is no code), refers to the institution where vouchers (plankton samples) are located. PLATFORM refers to the name of the ship or platform from which observations were made and can be found in the [s 3 platform.txt](#) file. The CRUISE code is the originator's cruise number. A REFERENCE code refers to the type of reference instrument that was used and can be found in the [s 40 ref instrument.txt](#) file. A PI code can be found in [primary_investigator_list.txt](#)

In the Version 2 format, the CRUISEINFO section can also have TS PROBE, GEAR, and METHODS information.

(2) STATION section:

The STATION section always consists of three rows and eleven columns. The column order is reserved, an example of which is as follows:

```
STATION  431
LAT DEG  LAT MIN  LAT SEC  LAT HEM  LONG DEG  LON MIN  LON SEC  LON HEM  MONTH  DAY  YEAR
23       2          N        60       W        10      25     1964
```

This example shows the first part of a record for Station 431 (originator's station number), with a latitude of 23° 2' N (no seconds), a longitude of 60° W (no minutes or seconds), and the date of October 25, 1964.

Row 1 holds the STATION label in the first column and the originator's station number.
 Row 2 holds the labels for the location/date information; and
 Row 3 holds the data for the labels in row 2.

(3) HEADERS section:

The HEADERS section can have a flexible number of rows in it, depending on how much metadata has been provided by the originator. It is five (5) columns wide.

In general:

- (1) Column 1 holds the label,
- (2) Column 2 holds the value, and
- (3) Column 3 holds the units.

The exceptions are ending locations (LAT END, LONG END) and times (TIME), where column 2 is degrees or hours, column 3 is minutes, column 4 is seconds, and column 5 is time zone or hemisphere.

A header section for station data might look like this:

	(1)	(2)	(3)	(4)	(5)
HEADERS					
TIME		12	14		UT
LAT END		23	50		N
LONG END		60	30		W
TS PROBE		CTD			
BOTTOM DEPTH		1000	m		
OXY METHOD		WINKLER			
WINDDIR		100	degrees		

The time is 12 hours, 14 minutes, and the time zone is UT.

If the dataset has been obtained from the OCL database and some parameters have been added to an already existing cruise, the HEADERS section might look like this:

	(1)	(2)	(3)	(4)	(5)
HEADERS					
TIME			1.7		GMT
ORIGINAL NODC ACCESSION#		6500000	OCLcode		
PLATFORM		1364	OCLcode		
OCL UNIQSTAT		436861	OCLcode		
BOTTOM DEPTH		5709	m		
WINDFOR		3	OCLcode		
WINDDIR		9	OCLcode		
WEATHER		-1	OCLcode		
LATITUDE SIG		4	OCLcode		
LONGITUDE SIG		3	OCLcode		
TIME SIG		2	OCLcode		

In this example, “OCLcode” in Column 3 indicates that the codes in Column 2 are OCL codes, which can be found in http://www.nodc.noaa.gov/OC5/WOD/wod_codes.html.

For the TIME parameter, minutes are entered as a decimal part of hour. It is calculated as follows:

$$42 \text{ min} / 60 \text{ min} = 0.7 \text{ hour}$$

$$1 \text{ hour} + 0.7 \text{ hour} = 1.7 \text{ hour}$$

The code for WINDFOR can be found in the windfor.txt file. The WINDDIR code can be found in the winwaved.txt file. The code for WEATHER can be found in one of two files: weather1.txt or weather2.txt.

The LATITUDE SIG, LONGITUDE SIG, and TIME SIG rows all indicate the number of significant figures to the right of the decimal point.

Another example of the HEADERS section for data may look like this:

	(1)	(2)	(3)	(4)	(5)
HEADERS					
LAT END		45	14.5		N
LONG END		163	45.7		E
TIME		9	55		local
TIME END		10	30		local
BOTTOM DEPTH		6170	m		
WEATHER		cloudy			
AIRTEMP		9.3	C		
WINDDIR		E	compass		
WINDSP		8	m/s		
BARPRESS		1019.7	mbar		
SEA		3	code		
SWELL		1	code		
VISIBILITY		7	code		

The code for SEA is found in seastate.txt. Since OCL does not store SWELL, there is no code table. However, if this variable was provided by originator, it will be entered. The code for VISIBILITY is found in visibil.txt.

A typical header section for a biological sample might look like this:

	(1)	(2)	(3)	(4)	(5)
HEADERS					
TIME		9	55		local
BOTTOM DEPTH		6170	m		
BIOTIME		11	12		local
BIOTIME END		11	20		local
GEAR		NORPAC			
MESH SIZE		0.33	mm		
TOW TYPE		V			
CHL METHOD		spectrometric			
FILTER TYPE		Reeve Angel gf/f 984H			

(4). DETAILS section:

This section can have any number of columns and rows beyond the three mandatory title rows (DETAILS, UNITS, and DECIMAL PLACES). The variable labels appear in the first row (DETAILS). The units that correspond to these labels appear in the second row (UNITS). The number of significant figures to the right of the decimal point appear in the third row (DECIMAL PLACES).

In the Version 1 format, there will not be a UNITS title row and the depth of observation will be found in the first column, a sample of which might look like this:

DETAILS	TEMP	SAL	...(other variables)	← variable name
DEPTH				
m	C	psu		← variable unit
DECIMAL PLACES	2	3		← number of figures to the right of the decimal point
0	27.48	36.182		
1				
10	27.5	36.188		
.	.	.		
.	.	.		
.	.	.		

In the Version 2 format, for depth-dependent data (e.g., temperature and salinity) and non-taxonomic data (e.g., production and biogeochemical fluxes), the depth of observation will be in the second column. A typical DETAILS section for station data in the Version 2 format might look like this:

DETAILS	DEPTH	TEMP	SAL	...(other variables)	← variable name
UNITS	m	C	psu		← variable unit
DECIMAL PLACES	0	2	3		← number of figures to the right of the decimal point
	0	27.48	36.182		
	1				
	10	27.5	36.188		
	.	.	.		
	.	.	.		
	.	.	.		

For taxonomic or integrated-depth observations, UPPER DEPTH and LOWER DEPTH are provided. A typical DETAILS section for a biological sample might look like this:

DETAILS	UPPER DEPTH	LOWER DEPTH	TAX COUNT	TAX PRESENT	TAX NAME
UNITS	m	m	#/ml	code	name
DECIMAL PLACES	0	0	0		
	0	0	10		Achnanthes sp.
	0	0	250		Asteromphalus
	0	0		abundant	Chaetoceros
	0	20		rare	Achnanthes sp.
	0	20	20		Asteromphalus
	0	20	40		Chaetoceros

Or the same data might be entered like this:

DETAILS	TAX CNT B0	TAX CNT B20	TAX PRS B0	TAX PRS B20	TAX NAME
UNITS	#/ml	#/ml	code	code	Name
DECIMAL	0	0	0	0	
	10			rare	Achnanthes sp.
	250	20			Asteromphalus
		40	abundant		Chaetoceros

Where, TAX CNT B0 and TAX CNT B20 stand for TAX COUNT at depth 0 or depth 20, respectively. TAX PRS B0 and TAX PRS B20 stand for TAX PRESENT at depth 0 or depth 20, respectively. The letter “B” in the TAX CNT B0, etc., can also be represented by the letter “Z”, i.e., TAX CNT Z0 etc.

Note: To the best of our ability, taxonomic names have been checked against the Integrated Taxonomic Information System (<http://www.itis.gov>)

A typical DETAILS section combining both depth-dependent and integrated-depth observations might look like this:

DETAILS	DEPTH	TEMP	PRIM PROD	UPPER DEPTH	LOWER DEPTH	PRIM PROD_INT
UNITS	m	C	mgC/m3/hr	m	m	mgC/m2/hr
DECIMAL	0	2	2	0	0	1
	0	22.48	1.57	0	80	23.2
	10	22.01	1.32			

A typical DETAILS section with only integrated-depth observations might look like this:

DETAILS	UPPER DEPTH	LOWER DEPTH	TAX COUNT	TAX NAME
UNITS	m	m	#/ml	
DECIMAL	0	0	0	
	0	0	608	Particulates 2.0-40.0um
	52	52	504	Particulates 2.0-40.0um
	0	0	83	Suspended solid
	12	12	44	Suspended solid
	33	33	55	Suspended solid

Should you require additional help or to report any problems, please contact OCL.help@noaa.gov.