# AZMP Data Assembly Documentation for BioChem Reload

Prepared by Gordana Lazin, October 11, 2017

Updated December 1, 2017

# Scope

This document describes the assembly of the data and metadata for 27 AZMP spring and fall cruises that occurred from 1999 to 2012 (Table 1). The assembly took place during 2015, 2016 and 2017 and was a part of the BioChem reload project. This document is intended for internal use in the datashop at the Bedford Institute of Oceanography and includes the details of the assembly process along with references to the scripts and key files.

|  |  |  |  |
| --- | --- | --- | --- |
| **Year** | **Regional Mission ID** | **Chief Scientist** | **Season** |
| 1999 | HUD99054 | Erica Head, Michel Mitchell | FALL |
| 2000 | HUD2000050 | Edward Horne, Michel Mitchell |
| 2001 | HUD2001061 | Edward Horne, Michel Mitchell |
| 2002 | HUD2002064 | Edward Horne, Michel Mitchell |
| 2003 | HUD2003067 | Edward Horne |
| 2003 | HUD2003078 | Erica Head |
| 2004 | HUD2004055 | Erica Head |
| 2005 | HUD2005055 | Erica Head |
| 2006 | HUD2006052 | Erica Head |
| 2007 | HUD2007045 | Erica Head |
| 2008 | HUD2008037 | Erica Head, Edward Horne |
| 2009 | HUD2009048 | Erica Head |
| 2010 | HUD2010049 | Erica Head |
| 2011 | HUD2011043 | Erica Head |
| 2012 | HUD2012042 | Edward Horne |
| 1999 | HUD99003 | Edward Horne | SPRING |
| 2000 | PAR2000002 | Edward Horne, Michel Mitchell |
| 2001 | HUD2001009 | Erica Head, Michel Mitchell |
| 2002 | **No cruise** |  |
| 2003 | HUD2003005 | Erica Head |
| 2004 | HUD2004009 | Edward Horne |
| 2005 | NED2005004 | Edward Horne |
| 2006 | HUD2006008 | Edward Horne |
| 2007 | HUD2007001 | Edward Horne |
| 2008 | HUD2008004 | Edward Horne |
| 2009 | HUD2009005 | Edward Horne |
| 2010 | HUD2010006 | Edward Horne |
| 2011 | HUD2011004 | Edward Horne |
| 2012 | **No cruise** |  |

Table 1: List of historical AZMP cruises for BioChem Reload.

# Data Assembly Process

The data preparation work was conducted in three phases:

1. Source data assembly, which involves locating, re-formatting, and correcting source data files, and in some cases re-creating or re-processing of the source data.
2. Assembly of BioChem load tables, which involves extracting necessary information from various data sources to construct metadata and data files in appropriate format for loading into BioChem database (creating BCS and BCD files respectively). In this step extensive effort was required for data cleaning.
3. Quality control and flagging that was performed on the metadata and data in the load tables and is based on the IML protocol for bottle data quality control.

## Source Data Assembly

Source data for AZMP cruises was gathered for each cruise and placed in the folder named 1\_REBOOT that was created in each individual cruise folder in the source directory and include:

1. **BiolSum files**, which contain nutrient, chlorophyll, and oxygen data.
2. **QAT files**, which contain CTD data at the depth that the bottle was closed.
3. **ODF files**, which contain CTD profiles and important metadata for each CTD cast.
4. **Electronic bridge log**, which contains information about events during cruise.
5. **Bottle data** files containing various parameters (HPLC, CHN etc.)
6. Bridge log book that was used to verify inconsistent metadata.
7. CTD dry deck sheets that were used to verify any inconsistencies.

The instruction for AZMP data assembly that was created for Inna Yashayaeva and Jay Bugden who were helping with data assembly can be found here:

[\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\Data\_by\_Year\_and\_Cruise\BioChem\_Data\_Preparation\_Hudson\_AZMP.docx](file:///\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\Data_by_Year_and_Cruise\BioChem_Data_Preparation_Hudson_AZMP.docx)

The assembly included following edits to the source data in excel format (BiolSums, Bridge Logs, and data files):

* Three worksheets were created in each file named “*parameter*\_FOR\_RELOAD”, “MAP”, and “README”.
* Original data was pasted to the “*parameter*\_FOR\_RELOAD” worksheet.
* Column names were changed in “*parameter*\_FOR\_RELOAD” to achieve consistent column naming.
* Mapping of the original columns to the new column names is documented in the “MAP” worksheet.
* Any required changes to the data were made in the “*parameter*\_FOR\_RELOAD” worksheet and were documented by highlighting the changed data, by attaching comments to the changed cells, and by describing changes in the “README” worksheet. The data in original worksheets stayed intact.

The naming conventions for the column are documented here:

[\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\Short\_Names\_BioChem.xlsx](file:///\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\Short_Names_BioChem.xlsx)

QAT file was created for the whole cruise by concatenating QAT files for the individual CTD casts using Python scripts and the columns were named according to the convention in the “Short\_Names\_BioChem.xlsx” file.

ODF metadata was created for each cruise by using Matlab script that reads ODF headers and writes out a list of CTD metadata for the cruise.

## Assembly of BioChem Load Tables

In order to assemble data to load tables for BioChem reload a toolbox was developed in R programing language that reads data from various sources, performs various checks, and outputs the data and metadata in a desired format (BCS and BCD files) as well as associated report file, plots, and maps. When inconsistency of any kind is encountered the scripts diagnose the issue, write out the message in the report file, and terminate the process. The issues that were identified have to be investigated, the source files needs to be corrected, and the scripts can be run again. The process is repeated until all inconsistencies are resolved.

During the assembly a large amount of inconsistencies were discovered for each mission and extensive corrections often had to be made. As a result, cleaning of source data for each mission required considerable amount of time. All changes to the source data are documented either in the excel files containing data or in a readme file in the 1\_REBOOT folder, and in the reload notes documents.

### BCS Table

BCS metadata table was assembled using R script “*bcs\_azmp1.r*”. The script performs following tasks:

-Reads a file containing a list of input files associated with each cruise:

[\\ent.dfo-mpo.ca\ATLShares\Science\BIODataSvc\SRC\BIOCHEMInventory\Data\_by\_Year\_and\_Cruise\Files\_for\_DIS\_header.csv](file:///\\ent.dfo-mpo.ca\ATLShares\Science\BIODataSvc\SRC\BIOCHEMInventory\Data_by_Year_and_Cruise\Files_for_DIS_header.csv)

-Creates a report file for logging all issues.

-Reads files containing metadata for one cruise (BiolSum , QAT file, ODF metadata, and Bridge Log), checks each one for internal inconsistencies, and converts dates and time in common formats using custom made functions (*check\_biolsum1.r, check\_ctd\_metadata.r, check\_qat1.r,* and *check\_bridgeLog.r*)

-Compares metadata fields from different sources (time, location, depth, station names etc.), performs metadata quality control (depth check, locations on land etc.), match bottle data with CTD data based on various metadata fields (CTD event number, sample ID, date/time, location, station name etc.) and produces comparison plots (for example CTD start date/time/location versus bridge log start date/time/location)

- Selects columns for BCS table, retrieves cruise information from OSC cruise database, outputs BCS file and associated report file, plots, and an interactive cruise track map in the html format.

The source of information for each BCS field is shown in Table 2. BCS files were uploaded to the Oracle staging table AZMP\_MISSIONS\_BCS on Gordana Lazin account and the aces was granted to Shelley Bond and Jay Bugden. It includes 27 cruises and 18962 records. BCS files were also placed in the 1\_REBOOT folder for each cruise.

|  |  |
| --- | --- |
| **Header** | **Data Source (Example Data)** |
| DIS\_SAMPLE\_KEY\_VALUE | Created by R script (18HU13004\_event\_SampleID) |
| MISSION\_DESCRIPTOR | OSCCRUISE database (18HU13004) |
| EVENT\_COLLECTOR\_EVENT\_ID | Bridge Log (3 digit eventID, 087) |
| EVENT\_COLLECTOR\_STN\_NAME | Bridge Log (HL\_02) |
| MISSION\_NAME | **Input by User**  (HUD2013004) |
| MISSION\_LEADER | OSCCRUISE database (DAVE HEBERT) |
| MISSION\_SDATE | OSCCRUISE database (04-Apr-2013, required data format) |
| MISSION\_EDATE | OSCCRUISE database (27-Apr-2013) |
| MISSION\_INSTITUTE | DFO BIO (hard coded) |
| MISSION\_PLATFORM | OSCCRUISE database (CCG Hudson) |
| **MISSION\_PROTOCOL** | From *Files\_for\_BCS\_header.xls* (AZMP) |
| MISSION\_GEOGRAPHIC\_REGION | OSCCRUISE database (SCOTIAN SHELF) |
| MISSION\_COLLECTOR\_COMMENT1 | Multiple legs, created by R code & OSCCRUISE info (This cruise has 2 legs) |
| MISSION\_COLLECTOR\_COMMENT2 | null |
| MISSION\_DATA\_MANAGER\_COMMENT | Maritimes BioChem Reload (hard coded) |
| EVENT\_SDATE | CTD metadata |
| EVENT\_EDATE | QAT last bottle |
| EVENT\_STIME | CTD metadata |
| EVENT\_ETIME | QAT last bottle |
| EVENT\_MIN\_LAT | min (QAT file lat and Bridge log start lat) |
| EVENT\_MAX\_LAT | max (QAT file lat and Bridge log start lat) |
| EVENT\_MIN\_LON | min (QAT file lon and Bridge log start lon) |
| EVENT\_MAX\_LON | max (QAT file lon and Bridge log start lon) |
| EVENT\_UTC\_OFFSET | 0 (hard coded) |
| EVENT\_COLLECTOR\_COMMENT1 | Comments from Bridge Log |
| EVENT\_COLLECTOR\_COMMENT2 | null |
| EVENT\_DATA\_MANAGER\_COMMENT | null |
| DIS\_HEADR\_GEAR\_SEQ | 90000019 for 10L Niskin bottle data,  90000065 for CTD only |
| DIS\_HEADR\_SDATE | QAT file (each bottle has different time) |
| DIS\_HEADR\_EDATE | QAT file, same as start date (SDATE) |
| DIS\_HEADR\_STIME | QAT file (each bottle has different time) |
| DIS\_HEADR\_ETIME | QAT file, same as start time (STIME) |
| DIS\_HEADR\_TIME\_QC\_CODE | 0 |
| DIS\_HEADR\_SLAT | QAT file |
| DIS\_HEADR\_ELAT | QAT file, same as start latitude |
| DIS\_HEADR\_SLON | QAT file |
| DIS\_HEADR\_ELON | QAT file, same as end latitude |
| DIS\_HEADR\_POSITION\_QC\_CODE | 0 |
| DIS\_HEADR\_START\_DEPTH | "Nominal Depth" from BiolSum or QAT pressure for CTD only |
| DIS\_HEADR\_END\_DEPTH | End depth = Start depth |
| DIS\_HEADR\_SOUNDING | Bridge Log |
| DIS\_HEADR\_COLLECTOR\_DEPLMT\_ID | null |
| DIS\_HEADR\_COLLECTOR\_SAMPLE\_ID | BiolSum or QAT for CTD only |
| DIS\_HEADR\_COLLECTOR | From *Files\_for\_BCS\_header.xls* (AZMP- JEFF SPRY) |
| DIS\_HEADR\_COLLECTOR\_COMMENT1 | End depth=Start depth. Start depth is nominal. (for bottle data). or Start depth is pressure from CTD QAT file (for CTD only). |
| DIS\_HEADR\_DATA\_MANAGER\_COMMENT | BioChem reload, QC performed using modified IML protocols. No bottle data (for CTD only) + comments from BiolSum |
| DIS\_HEADR\_RESPONSIBLE\_GROUP | From *Files\_for\_BCS\_header.xls* (AZMP) |
| DIS\_HEADR\_SHARED\_DATA | null |
| CREATED\_BY | Hard coded (Gordana Lazin) |
| CREATED\_DATE | Created by R script (22-Nov-2015) |
| DATA\_CENTER\_CODE | 20 (hard coded in R script) |
| PROCESS\_FLAG | NR |
| BATCH\_SEQ | null |

Table 2: Source of information used to create BCS metadata table for AZMP cruises.

### BCS Issues Resolved

There were a large number of issues resolved in the metadata. For example CTD data was merged to the water sample data based on the event number and sample ID. It was discovered that in many missions CTD data was numbered by CTD cast number, instead by the event number so corrections have to be made in original CTD data (ODF headers and QAT files) before proceeding to the further assembly steps. Another common example is that sample IDs were typed incorrectly in QAT file so merging CTD data with water sample data based on the sample ID was not possible. Large amount of the information in the source data was entered manually so there were many typos and inappropriate information entered in the wrong fields (for example mixture of station names and CTD cast numbers entered in the CTD event number field). In some cases QAT files had local time instead of UTC, and data for several cruises did not account for the daylight saving time change. It was also found that “nominal depth” used for bottles in some cases greatly deviated from the pressure.  For example, a bottle fired at 2.9 m and the one fired at 17 m can be both associated with 10 m nominal depth. In consultation to the AZMP team it was decided to still use nominal depth for the bottle depth even though the real depth reflected by the pressure can be quite different. The issues resolved are documented in the notes in the “reload\_notes” directory found here: [\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\Documents\reload\_notes](file:///\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\Documents\reload_notes)

*azmp\_processing\_notes.docx*

*BioChem Reload Notes.docx*

*ODF data summary for AZMP cruises.docx*

*RE depth in biochem.msg* (or.txt) in the “QC” directory

### BCD Table

BCD data table was assembled using R script “bcd\_azmp1.r”. The script performs following tasks:

- reads the data files (QAT files for CTD data, BiolSum, HPLC, and CHN files) using list of the input file names contained in the file:

[\\ent.dfo-mpo.ca\ATLShares\Science\BIODataSvc\SRC\BIOCHEMInventory\Data\_by\_Year\_and\_Cruise\Files\_for\_DIS\_header.csv](file:///\\ent.dfo-mpo.ca\ATLShares\Science\BIODataSvc\SRC\BIOCHEMInventory\Data_by_Year_and_Cruise\Files_for_DIS_header.csv)

- maps the variables in the source files to the proper BioChem Data Type Method and assigns appropriate sequences.

- compares Sample IDs found in BCS file to the ones in the data files and reports the differences. Only the samples found in BCS were included in BCD.

- converts data from wide format to the long format required by BioChem.

- creates BCD table.

The data types and associated source files are shown in Table 3, and the data count in the BCD tables for each cruise are shown in Tables 4 and 5.

|  |  |  |
| --- | --- | --- |
| **Parameter (method)** | **Source File** | **Replicates** |
| Chl\_a\_Holm-Hansen\_F | BiolSum | single |
| Phaeo\_Holm-HansenF | BiolSum | single |
| NO2NO3\_Tech\_F | BiolSum | single |
| PO4\_Tech\_F | BiolSum | single |
| SiO4\_Tech\_F | BiolSum | single |
| NH3\_Tech\_F | BiolSum | single |
| NO2\_Tech\_F | BiolSum | single |
| O2\_Electrode\_mll | BiolSum | single |
| O2\_Winkler | BiolSum | single |
| Salinity\_Sal\_PSS | BiolSum | single |
| CO2 | BiolSum | single |
| HPLC | HPLC files | single |
| POC\_CHN\_mg/m3 | CHN files | duplicates |
| PON\_CHN\_mg/m3 | CHN files | duplicates |
| Temp\_CTD\_1968 | QAT files | single |
| Salinity\_CTD | QAT files | single |
| conductivity\_CTD | QAT files | single |
| Pressure | QAT files | single |
| O2\_CTD\_mLL | QAT files | single |
| Chl\_Fluor\_Voltage | QAT files | single |
| pH\_CTD\_nocal | QAT files | single |

Table 3: Data types included in the AZMP BCD table and the associated source files.

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | HUD99003 | PAR2000002 | HUD2001009 | HUD2003005 | HUD2004009 | NED2005004 | HUD2006008 | HUD2007001 | HUD2008004 | HUD2009005 | HUD2010006 | HUD2011004 |
| Chl\_a\_Holm-Hansen\_F | 327 | 594 | 598 | 220 | 642 | 312 | 443 | 564 | 508 | 499 | 646 | 604 |
| Phaeo\_Holm-HansenF | 327 | 594 | 598 | 220 | 642 | 312 | 443 | 564 | 508 | 499 | 646 | 604 |
| NO2NO3\_Tech\_F | 417 | 726 | 724 | 269 | 822 | 339 | 680 | 785 | 723 | 698 | 931 | 875 |
| PO4\_Tech\_F | 417 | 726 | 724 | 269 | 822 | 339 | 680 | 785 | 723 | 698 | 931 | 875 |
| SiO4\_Tech\_F | 417 | 726 | 724 | 269 | 822 | 339 | 680 | 785 | 723 | 698 | 931 | 875 |
| NH3\_Tech\_F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 698 | 931 | 875 |
| NO2\_Tech\_F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 698 | 930 | 875 |
| O2\_Electrode\_mll | 104 | 0 | 136 | 5 | 178 | 82 | 135 | 159 | 150 | 133 | 164 | 178 |
| O2\_Winkler | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 139 |
| Salinity\_Sal\_PSS | 0 | 0 | 0 | 32 | 0 | 0 | 57 | 72 | 141 | 142 | 178 | 172 |
| CO2 | 58 | 68 | 61 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HPLC | 67 | 156 | 175 | 48 | 72 | 34 | 36 | 62 | 46 | 55 | 71 | 69 |
| POC\_CHN\_mg/m3 | 78 | 126 | 223 | 56 | 142 | 68 | 72 | 122 | 92 | 112 | 142 | 138 |
| PON\_CHN\_mg/m3 | 78 | 123 | 223 | 56 | 142 | 68 | 72 | 122 | 92 | 112 | 142 | 138 |
| Temp\_CTD\_1968 | 529 | 797 | 831 | 282 | 858 | 414 | 722 | 822 | 759 | 895 | 1471 | 1025 |
| Salinity\_CTD | 529 | 797 | 831 | 282 | 858 | 414 | 722 | 822 | 759 | 895 | 1471 | 1025 |
| conductivity\_CTD | 529 | 797 | 831 | 282 | 858 | 414 | 722 | 822 | 759 | 895 | 1470 | 1025 |
| Pressure | 529 | 797 | 831 | 282 | 858 | 414 | 722 | 822 | 759 | 895 | 1471 | 1025 |
| O2\_CTD\_mLL | 529 | 797 | 831 | 282 | 858 | 414 | 722 | 822 | 759 | 895 | 1471 | 1025 |
| Chl\_Fluor\_Voltage | 0 | 797 | 0 | 282 | 858 | 414 | 722 | 822 | 759 | 895 | 1470 | 1025 |
| pH\_CTD\_nocal | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 895 | 1470 | 0 |

Table 4: Number of records in the BCD table for the spring AZMP missions.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | HUD99054 | HUD2000050 | HUD2001061 | HUD2002064 | HUD2003067 | HUD2003078 | HUD2004055 | HUD2005055 | HUD2006052 | HUD2007045 | HUD2008037 | HUD2009048 | HUD2010049 | HUD2011043 | HUD2012042 |
| Chl\_a\_Holm-Hansen\_F | 493 | 372 | 576 | 394 | 404 | 309 | 293 | 307 | 390 | 345 | 484 | 563 | 69 | 529 | 713 |
| Phaeo\_Holm-HansenF | 493 | 372 | 576 | 394 | 404 | 309 | 293 | 307 | 390 | 345 | 484 | 563 | 69 | 529 | 713 |
| NO2NO3\_Tech\_F | 601 | 579 | 729 | 488 | 525 | 452 | 355 | 354 | 480 | 492 | 716 | 813 | 205 | 773 | 1031 |
| PO4\_Tech\_F | 601 | 579 | 729 | 488 | 525 | 452 | 355 | 354 | 480 | 492 | 716 | 813 | 205 | 773 | 981 |
| SiO4\_Tech\_F | 601 | 579 | 729 | 488 | 525 | 452 | 355 | 354 | 480 | 492 | 716 | 813 | 205 | 773 | 1032 |
| NH3\_Tech\_F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 716 | 813 | 205 | 773 | 1032 |
| NO2\_Tech\_F | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 716 | 813 | 205 | 420 | 1032 |
| O2\_Electrode\_mll | 133 | 0 | 140 | 112 | 99 | 64 | 75 | 82 | 105 | 118 | 223 | 160 | 30 | 71 | 207 |
| O2\_Winkler | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 310 | 486 |
| Salinity\_Sal\_PSS | 0 | 0 | 0 | 0 | 0 | 0 | 31 | 62 | 95 | 162 | 247 | 244 | 163 | 212 | 194 |
| CO2 | 68 | 61 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| HPLC | 137 | 149 | 164 | 76 | 60 | 46 | 32 | 34 | 44 | 36 | 49 | 79 | 9 | 60 | 79 |
| POC\_CHN\_mg/m3 | 110 | 134 | 218 | 139 | 89 | 89 | 64 | 68 | 88 | 72 | 98 | 124 | 20 | 119 | 158 |
| PON\_CHN\_mg/m3 | 110 | 136 | 218 | 139 | 89 | 89 | 64 | 68 | 88 | 72 | 98 | 124 | 20 | 119 | 158 |
| Temp\_CTD\_1968 | 702 | 648 | 738 | 564 | 539 | 490 | 360 | 401 | 533 | 516 | 825 | 1046 | 219 | 800 | 1185 |
| Salinity\_CTD | 702 | 648 | 738 | 564 | 539 | 490 | 360 | 401 | 533 | 516 | 815 | 1046 | 219 | 800 | 1185 |
| conductivity\_CTD | 702 | 648 | 738 | 564 | 539 | 490 | 360 | 401 | 533 | 516 | 825 | 1046 | 219 | 800 | 1185 |
| Pressure | 702 | 648 | 738 | 564 | 539 | 490 | 360 | 401 | 533 | 516 | 825 | 1046 | 219 | 800 | 1185 |
| O2\_CTD\_mLL | 702 | 648 | 738 | 564 | 539 | 490 | 360 | 401 | 533 | 516 | 201 | 1046 | 219 | 800 | 1185 |
| Chl\_Fluor\_Voltage | 702 | 0 | 0 | 396 | 539 | 490 | 360 | 401 | 533 | 516 | 201 | 1046 | 219 | 800 | 1185 |
| pH\_CTD\_nocal | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1046 | 0 | 0 | 631 |

Table 5: Number of records in the BCD table for the fall AZMP missions.

## Quality Control

Initially the data in the BCD table were flagged as erroneous only to the extent that would allow loading BCD files into BioChem staging tables. The flag number 4 (value seems erroneous) was assigned to the following cases:

* Temperature outside range -2.5C to 35°C (global and regional range from IML protocol)
* Salinity outside range of 0 to 50 PSU (global range)
* Oxygen outside range of 0 to 11 ml/l (global range from IML protocol)
* Negative values for all other parameters

In the second round of flagging IML quality control Matlab scripts were applied to the data in the BCD table which resulted in the temperature, salinity, oxygen, and nutrient data (nitrate, phosphate and silicae) flagged in accordance to the IML protocols. IML Matlab QC toolbox was modified to include temperature and salinity climatology from Scotian Shelf (Petrie at al., 1996) and nutrient climatology for Scotian Shelf (Lazin et al. 2014, unpublished). Temperature and salinity regional ranges were also adjusted in the scripts (Matlab function B\_stage\_Q\_ini.m) so they are more appropriate for the Scotian Shelf.

Adoption of IML Matlab toolbox to our dataset required several steps of adjustments. Since the toolbox required very specific input data format, R scripts were developed to convert BCD table into formats that can be ingested by the IML scripts. For each cruise two input files were created, one containing CTD data and the other containing bottle data (for example 18HU11043\_IML\_format\_ctd.txt and 18HU11043\_IML\_format.txt) and those files were run separately through the toolbox. The reason was that IML scripts are not checking CTD data, but only data marked as “labo”. To QC CTD data, temperature, salinity and oxygen were made to look as “labo” variables in the input files. The output of the IML script includes several different files containing flags and the explanation for flagging. In the last step the flags from the IML output files were transferred to the BCD files using R scripts, and finally the \*BCD\_flagged.csv files were created for each cruise.

Since large amount of data from 27 cruises had to be flagged, the IML toolbox was further modified so the cruises can be run in the batch mode. As a consequence interactive profile plotting intended for the visual inspection of the profiles was turned off. As a replacement for visual inspection, R script was developed that plots only flagged profiles so that suspect points can be further reviewed (Figure 1). The profiles that were not flagged by IML script were not visually inspected. All QC was performed in the IML\_QC directory of the working folder, which contains all input and output files. Modified IML toolbox is on SVN and is named IML\_QC\_2017. The “*cruise*\_BCD\_flagged.csv” files are presently located in the QC folders of the individual cruise folders.

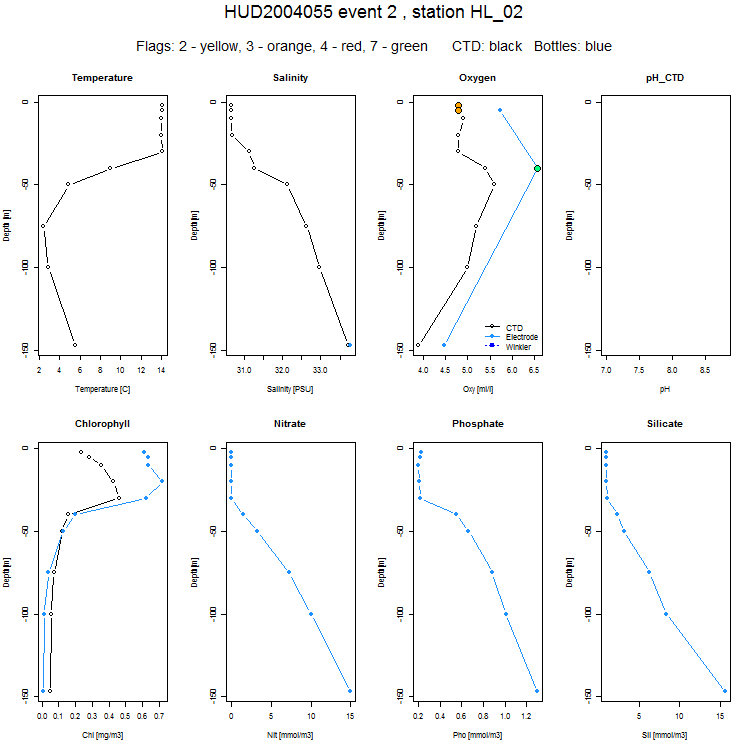


Figure 1: An example of the profiles flagged by IML QC toolbox. Only flagged profiles are plotted for further evaluation.

The last step in flagging involves flag review by the data provider. Plots of the flagged profiles along with the table for tracking flag changes (*\*\_change\_flag\_log.xslx*) were sent to the data provider (Andrew Cogswell) for the review. At this time the review of the flags is not completed and the BCD staging table does not include IML flags.

### QC update, December 1, 2017

The flagging procedure outlined above produced large number of flags (more than 5000) and majority of those flags were assigned to CTD data, which made flag review extremely overwhelming . The QC team met on November 1st, 2017 to discuss the issue (Andrew Cogswell, Shelley Bond, Gordana Lazin, and Jeff Jackson). To simplify BioChem flagging and flag review for historical and recent bottle data we made following decisions:

1. CTD data (QAT data) in biochem will not be flagged. That includes Temperature, Salinity, CTD Oxygen, and pH.  ODF database with CTD data will be eventually linked to BioChem and will contain correct and up to date QAT data. Flagged data will include nutrients, chlorophyll, and Winkler oxygen.
2. Oxygen from electrode will not be flagged. It is known that electrode oxygen data is not accurate anyway.

In the current version of flagged data the flags produced by IML scripts for CTD parameters were removed. The flag #4 (invalid) is still assigned to the CTD parameters that are out of range.

Original version of flagged data (\*\_BCD\_flagged\_cdt.csv) and associated plots of the flagged profiles were placed in a subfolder of the 1\_REBOOT folder for each cruise named:

\1\_REBOOT\QC\**old\_ctd\_flagged\_profiles**

The new version of flagged data, without CTD flags (\*\_BCD\_flagged.csv) can be found in the QC folder and all associated flagged profiles and the flag report in the folder named:

\1\_REBOOT\QC\**flagged\_profiles**

The new flagging practice reduced the number of flags by about 80%.

The summary of the flags can be found in the following file:

[\\ent.dfo-mpo.ca\ATLShares\Science\BIODataSvc\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\QC\_notes](file:///\\ent.dfo-mpo.ca\ATLShares\Science\BIODataSvc\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\QC_notes)

# Scripts

All tasks were performed using R scripts that were developed as part of the reboot toolbox. The main scripts used in AZMP data assembly, Oracle upload, and flagging are listed in Table 11. Thera are many more scripts required to execute main scripts and they are all listed as required in the beginning of the main scripts. All scripts written for reboot project can be found in the working directory and on SVN in \ODS\_Toolbox\R\R\_Code\Biochem\_Functions.

|  |  |
| --- | --- |
| Step | Action / Script Name |
| 1. Create BCS file | *bcs\_azmp1.r* |
| 1. Create BCD file | *bcd\_azmp1.r* |
| 1. Upload to Oracle staging table | *write\_biochem\_table\_RODBC.r* |
| 1. Convert BCD to IML format | *BCD2IML\_format.r* |
| 1. Run IML format through Matlab QC scripts | IML QC Matlab scripts: *B\_batch\_BIO.m* |
| 1. Export IML flags to BCD file | *get\_flags\_IML2BCD.r*  provides summary of flags |
| 1. Plot flagged profiles and reports for flags | *plot\_flagged\_profiles1.r* |
| 1. Evaluate flags | Look at the plots and reports – flag review by Andrew |
| 1. Keep track of changed flags | In the file *HUDYY0##\_change\_flag\_log.xlsx* |
| 1. Insert modified flags to BCD | Write a script, or insert by hand if not many |
| 1. Reload flagged BCD to staging table | *write\_biochem\_table\_RODBC.r* needs to be modified |

Table 6: Steps in AZMP data assembly along with the associated scripts. Shaded steps are not completed.

# Comparison of Reboot dataset to the current BioChem records

## BCS Metadata Comparison

SQL script was developed by Shelley Bond that compares metadata in BCS staging table to the AZMP metadata presently residing in BioChem. The discrepancies were reviewed by Jay Bugden and were all explained and resolved. They were sometimes due to the different time stamps (24 hour clock notation in BCD vs 12 hour clock notation in BioChem), daylight saving time adjustments, or small time difference caused by the time acquired from different data sources (bridge log time vs CTD time). The SQL script and the results of the metadata comparison can be found here: [\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\biochem\_comparison\BCS\BC\_reboot\_EVdatetime\_BCS\_DD\_JB\_GL.xlsx](file:///\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\biochem_comparison\BCS\BC_reboot_EVdatetime_BCS_DD_JB_GL.xlsx)

## BCD Data Comparison

R scripts were developed to compare records in BCD staging table to the records presently residing in BioChem. The results of comparison are summarized in the two types of files that can be found in the following folder:

[\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\biochem\_comparison\BCD](file:///\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\biochem_comparison\BCD)

The file named “*cruise*\_counts\_comparison.csv” contains the number of records for each data type in BCD table alongside the number of records in the BioChem for each particular cruise. Note that the same parameter could have been mapped to different data type in the original load so the counts for each data type are shown separately (for example chlorophyll-a was mapped to “Chl\_a” data type in the original BioChem load while in the BCD table it was mapped to the “Chl\_a\_Holm-Hansen\_F”).

The files named “*cruise*\_comparison\_report.txt” contain counts comparison for each parameter, regardless of the mapping, and report the sample IDs that were found in one table but not in the other. This file can be used to further investigate the causes of differences between BCD staging table and the records in the BioChem tables and identify missing data. Some of the differences are due to the fact that flow-through samples were not included in the BCD reboot dataset while they were originally loaded to BioChem.

# Documentation

All documentation created during the project can be found in the following folder and is organized in the subfolders:

[\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\1\_REBOOT\_Gordana\Biochem\_reload\Documents](file:///\\dcnsbiona01a\BIODataSvc\SRC\BIOCHEMInventory\1_REBOOT_Gordana\Biochem_reload\Documents)

The documentation includes:

**Progress reports,** containing useful details of the BioChem Reload project at several stages. Examples of assembly process for AZMP are described in *BioChem\_Reload\_progress\_GL\_10Mar2017.docx*.

**Quality Control** documentation, with important files listed below:

-AZMP data assembly and QC flow diagram: *Bottle\_data\_QC\_flow.pptx*

-IML toolbox documentation, sent from IML: *BOTL\_ODF\_Qualite\_en.pdf, IML\_description\_BTL.docx*

-Flag review instructions: *Biochem\_flagging\_instructions.docx*

-Current flagging flow diagram for AZMP: *flagging\_flow.vsd (visio file)*

-Emails regarding IML QC Matlab toolbox exchanged with Laure Devine and Caroline Lafleur that contain many important information about QC practices at IML and about toolbox troubleshooting. Emails are saved as outlook \*.msg files and also as .txt files: *RE Matlab bottle data QC scripts.msg (.txt)*

**Tech Notes** contain various files of particular technical nature, one of which are:

*ODF\_Troubleshooting.docx* (procedure for replacing cast number with event number in ODF files)

**Reload Notes**, contain files with notes that were taken during various phases of reload project.

# Outstanding Work

-Complete flag review by data provider and supply *\*\_change\_flag\_log.xslx* to the datashop

- Insert modified flags to \*BCD\_flagged.csv files (write R script, or insert flags manually if not too many)

- update Oracle staging table with the flags.

- consider flagging other parameters by introducing acceptable ranges. So far flagging included temperature, salinity, oxygen, and nutrients (nitrate, phosphate and silicate) as per IML protocol.

- review comparison between records in BCD tables and the data currently in BioChem and identify if anything is missing.

- replace CTD cast number with event number in ODF file headers and file names and/or upload them to the cdt archive. The files relating CTD cast number to event number were sent to Jeff Jackson. The cruises for which ODF files refer to the CTD cast number are following:

HUD1999-003

HUD1999-054 – ODF fixed by Inna, not in the archive

HUD2000-050

PAR2000-002

HUD2001-009

HUD2001-061– ODF fixed by Inna, not in the archive

HUD2002-064– ODF fixed by Inna, not in the archive

HUD2003-005

HUD2003-067 – fixed ODF are in the archive

HUD2003-078

HUD2004-009

NED2005-004

# Appendix: Discussions/Decisions

Below are few emails summarizing discussions/decisions made during the BioChem reload project.

**From:** Lazin, Gordana   
**Sent:** February-23-16 1:19 PM  
**To:** Bond, Shelley A; Benjamin, Robert; Cogswell, Andrew  
**Subject:** **BioChem: CTD casts without bottle data**

Hi All,

Here is the summary of the discussion that we had this morning regarding historical bottle and CTD data for loading to BioChem:

1. If there is bottle data and QAT file they will be both loaded with the GEAR\_TYPE: Bottle, and GEAR\_SEQ: 90000019
2. If there is no bottle data for a CTD cast, QAT file with CTD data will be loaded anyway, with the GEAR\_TYPE: CTD and GEAR\_SEQ: 90000065 (for SBE 911)

2a. Test casts will be loaded using sample ID from QAT file (that are often 1,2,3,4 etc or 999999). Sample ID will not be unique and there will be many single digit or 999999 ID’s in BioChem.

2b. If the samples were collected for some other party and we don’t have the data, QAT files will be loaded using sample ID from QAT file (if they exist) or from the CTD dry deck sheets.

2c. if the bottle misfired, there will be no bottle data or QAT data. The sample ID assigned originally to the misfired bottle will not be loaded. I will keep notes about those samples.

Please correct me if I got anything wrong. Thanks, Gordana

**From:** Lazin, Gordana   
**Sent:** 2016–September-20 4:59 PM  
**To:** Johnson, Catherine; Cogswell, Andrew; Devred, Emmanuel; Bond, Shelley A; Casault, Benoit  
**Cc:** Benjamin, Robert  
**Subject:** **depth in biochem**

Hi All,

As a part of QC of bottle data for BioChem reload I am comparing “nominal depth” at which bottles are supposed to be fired and “pressure” at which the bottles are actually fired.  In many cases there are discrepancies between nominal depth and actual pressure.  For example, a bottle fired at 2.9 m and the one fired at 17 m can be both associated with 10 m nominal depth.

Current practice for BioChem is to record nominal depth in DEPTH column and include pressure in PRESSURE column, so they are both available in BioChem.

My questions is: If there is a large discrepancy between nominal depth and pressure, should we be changing nominal depth as part of QC to better reflect actual depth?

If so, what would be considered “large discrepancy” for science purposes?

If not, we can continue with current practice but make sure that the users are aware of the “DEPTH” column meaning.

I am attaching an example of a cruise to show both nominal depths (depth\_bs column) and associated pressure (pressure\_qat, last column).

Thanks for your input, Gordana

**From:** Cogswell, Andrew   
**Sent:** September-29-16 2:20 PM  
**To:** Lazin, Gordana; Johnson, Catherine; Devred, Emmanuel; Bond, Shelley A; Casault, Benoit  
**Cc:** Benjamin, Robert  
**Subject:** **RE: depth in biochem**

Hi Gordana,

Yes, we see this very often to varying degrees.  The way these data are entered into BioChem is useful despite the discrepancy between nominal and actual bottle depths.  The nominal depths allow for an efficient means of extracting data from the database based on defined criteria, but the pressure values would allow the analyst to further refine/filter these data for final analysis.  For some analyses a 5 m or 3% error between nominal and actual could be totally fine, and for other analyses this level of discrepancy would be unacceptable.

As you say below:  “ we can continue with current practice but make sure that the users are aware of the “DEPTH” column meaning.”

The longer term goal is to separate the CTD sensor data (ODF Archive) from the BioChem discrete data (laboratory analysis).  The meta-data in BioChem would represent our sampling intent (nominal depths) while the associated CTD sensor data in the ODF/QAT file archives would represent sampling reality.  For now however, my personal opinion is to leave it as it is and allow the analyst to refine these data to suit their needs (with the further description you suggest above).

Andrew

**From:** Lazin, Gordana   
**Sent:** May-15-17 3:50 PM  
**To:** Cogswell, Andrew; Bond, Shelley A  
**Subject:** **BioChem QC Flags meeting**

Hi Shelley and Andrew,

Attached are the documents that I brought to the meeting. Here are the actions:

1.  Gordana will send Andrew files for one cruise so he can try to change flags manually, and evaluate difficulties with this method.

2.  Jeff Jackson will estimate the effort required to create an app for data flagging in R or Python, and come up with specifications.

3.  Gordana will find out about GUIs  and interactive plots at the Python course this week.

Thanks, Gordana

**From:** Cogswell, Andrew   
**Sent:** June-26-17 11:20 AM  
**To:** Lazin, Gordana  
**Cc:** Bond, Shelley A; Cogswell, Andrew  
**Subject:** **RE: Test QC cruise flagging**

Hi Gordana (a note so both of us can remember this conversation),

As we’ve discussed now at length, there appears to be an offset for all variables between the odf trace and the QAT/deck sheet data for this mission (99054).  All of the oxygen/salinity/fluorescence data should be considered uncalibrated CTD data.  There is no feedback mechanism (even now) to port calibrated data back into BioChem even though more recent AZMP missions now have regularly calibrated oxygen and salinity values in the reprocessed ODF and QAT files.

Shelley had suggested we hold a meeting sometime in September to discuss this further and make a decision about the level of QC we should apply on CTD related data in BioChem in light of the fact it may be removed in the near future upon completion of and linking to the ODF/QAT archive.  So, part of are discussion with Laure and others is what they are currently doing and whether we should continue with their model or develop a new one.

Andrew

**From:** Lazin, Gordana   
**Sent:** November-01-17 4:14 PM  
**To:** Cogswell, Andrew; Bond, Shelley A; Jackson, Jeffrey  
**Cc:** Lazin, Gordana  
**Subject:** **RE: BioChem QC meeting**

Decisions/actions from BioChem QC meeting, Nov 1st, 2017

To simplify BioChem flagging and flag review for historical  and recent bottle data we made following decisions:

1. CTD data (QAT data) in biochem will not be flagged. That includes Temperature, Salinity, CTD Oxygen, and pH.  ODF database with CTD data will be eventually linked to BioChem and will contain correct and up to date QAT data. Flagged data will include nutrients, chlorophyll, and Winkler oxygen.
2. Oxygen from electrode will not be flagged. It is known that electrode oxygen data is not accurate anyway.

Actions:

1. Gordana will re-set CTD and electrode oxy QC flags back to zero (not flagged) in the BCD files. That will eliminate about 80% of the flags (about 2500 flags) and the review will be required only for the remaining ~500 flags.
2. Andrew will review remaining flags when he finds time.
3. Gordana will transfer flagging procedures/scripts to Jeff Jackson, so flagging can continue for the recent cruises. Jeff will then teach Jay how to flag BCD data with IML scripts.
4. If some flags get modified during review someone will have to update BCD file and insert modified flags. After that BCD for AZMP historical data will be ready to load.

Other:

CTD metadata in the excel files in CTD archive are not up to date and will have to be removed from the archive to avoid confusion, since some of the ODF data file names have changed (cast numbers were replaced with event numbers).

I think this is all. Let me know if I missed or got anything wrong.

Thanks for coming to the meeting, Gordana

**From:** Lazin, Gordana   
**Sent:** September-12-17 1:44 PM  
**To:** Jackson, Jeffrey; Bond, Shelley A; Hackett, Jennifer  
**Subject:** **ODF files in the archive**

Hello,

I was checking the ‘ctd’ folder in the archive and noticed that for number of AZMP cruises we still have ODF files labeled by the CTD cast number instead by the event number. Inna fixed ODF files for some of those cruises and I was wondering if the rest of the cruises were fixed.

Here is the list of the cruises with that issues, and the corrections that I am aware of:

HUD1999-003

HUD1999-054 – ODF fixed by Inna, not in the archive

HUD2000-050

PAR2000-002

HUD2001-009

HUD2001-061– ODF fixed by Inna, not in the archive

HUD2002-064– ODF fixed by Inna, not in the archive

HUD2003-005

HUD2003-067 – fixed ODF are in the archive

HUD2003-078

HUD2004-009

NED2005-004

I am not sure if scientists are pulling data from that archive but if so there will be lot of confusion when trying to match ODF data labeled by the CTD cast number to the rest of the data that go by the event number.

If ODF’s are not fixed yet we should at least warn Andrew and the rest of the users. I had corrected all QAT files so they all refer to the event number now. Thanks, Gordana