BioChem

Content Audit

Version 2

13/11/2014

Shelley Bond, Maritimes

Table of Contents

[Table of Tables 2](#_Toc403657283)

[Abstract 4](#_Toc403657284)

[Background 6](#_Toc403657285)

[Standard Practices 7](#_Toc403657286)

[The Data 8](#_Toc403657287)

[Raw Data Archival Issues 8](#_Toc403657288)

[BCMissions 8](#_Toc403657289)

[Protocol, In the Wrong Place? 9](#_Toc403657290)

[Incorrect Data Center Codes 9](#_Toc403657291)

[Duplicated Mission Records 10](#_Toc403657292)

[Trailing White Spaces in Name and Descriptor Fields 11](#_Toc403657293)

[Incorrect Start and End Dates 11](#_Toc403657294)

[Data Loaded to Wrong Mission 12](#_Toc403657295)

[BCComments 12](#_Toc403657296)

[BCEvents 13](#_Toc403657297)

[Duplicate Events 13](#_Toc403657298)

[Dates Outside of Mission Dates 13](#_Toc403657299)

[BCDiscreteHedrs 14](#_Toc403657300)

[Duplicate Discrete Headers 14](#_Toc403657301)

[Dates Outside of Event Dates 14](#_Toc403657302)

[When Depth is not a Depth 15](#_Toc403657303)

[When Gear is not a Gear 15](#_Toc403657304)

[Questionable Positions 16](#_Toc403657305)

[BCDiscreteDtails 16](#_Toc403657306)

[Duplicated Discrete Data 17](#_Toc403657307)

[Dissolved Oxygen Conversion Issues 17](#_Toc403657308)

[Data of Unknown Origin 19](#_Toc403657309)

[Nulls Set to Zero 19](#_Toc403657310)

[Negative Values Set to Zero 19](#_Toc403657311)

[Zeroes Stripped Out of Data 20](#_Toc403657312)

[Previous Data Values Assigned to Nulls 20](#_Toc403657313)

[Not All Provided Data Loaded 21](#_Toc403657314)

[Repeated Values for Different Samples 21](#_Toc403657315)

[BCPlanktnHedrs 22](#_Toc403657316)

[Duplicate Plankton Headers 22](#_Toc403657317)

[Plankton Dates Outside of Event Dates 23](#_Toc403657318)

[Nominal Positions 23](#_Toc403657319)

[Questionable Positions 23](#_Toc403657320)

[BCPlanktnGenerals 24](#_Toc403657321)

[Duplicate Plankton Data 24](#_Toc403657322)

[When Count is Not a Count 24](#_Toc403657323)

[The Lookup Tables 26](#_Toc403657324)

[Lack of Lookup Tables 26](#_Toc403657325)

[Lookup Table Housekeeping 26](#_Toc403657326)

[BCNatnlTaxonCodes 27](#_Toc403657327)

[Content with No Associated Data 27](#_Toc403657328)

[Authorities Which Aren’t 27](#_Toc403657329)

[Duplicated Records 28](#_Toc403657330)

[Possible Miss-mappings to AphiaId and TSN 29](#_Toc403657331)

[The Corrections 29](#_Toc403657332)

# Table of Tables

[Table 1 Collector station name example 6](#_Toc403656350)

[Table 2 Duplicate missions example 9](#_Toc403656351)

[Table 3 HUD2006019 missions corrected 10](#_Toc403656352)

[Table 4 BCComments and station info 11](#_Toc403656353)

[Table 5 Duplicate events example 12](#_Toc403656354)

[Table 6 Event outside of mission dates example 13](#_Toc403656355)

[Table 7 Duplicate discrete header examples 13](#_Toc403656356)

[Table 8 Discrete header dates outside of event dates example 14](#_Toc403656357)

[Table 9 Sediment data and depths example 14](#_Toc403656358)

[Table 10 Ice gears, depths, and parameter example 15](#_Toc403656359)

[Table 11 Duplicated discrete data example 16](#_Toc403656360)

[Table 12 Discrete data converted and not converted duplicate example 16](#_Toc403656361)

[Table 13 Dissolved oxygen conversions 17](#_Toc403656362)

[Table 14 Dissolved oxygen conversion issues 18](#_Toc403656363)

[Table 15 Stripped out zeroes 19](#_Toc403656364)

[Table 16 Example of ‘recycled’ data values 20](#_Toc403656365)

[Table 17 Examples of missing sample data 20](#_Toc403656366)

[Table 18 Repeated values 21](#_Toc403656367)

[Table 19 Duplicate plankton header example 22](#_Toc403656368)

[Table 20 Plankton header date outside of event dates example 22](#_Toc403656369)

[Table 21 Duplicate plankton example 23](#_Toc403656370)

[Table 22 Platform variation example 25](#_Toc403656371)

[Table 23 Confusing lookup table example 26](#_Toc403656372)

[Table 24 Invalid lookup table content 26](#_Toc403656373)

[Table 25 Duplicate species examples 27](#_Toc403656374)

[Table 26 Using lookup table to track source example 27](#_Toc403656375)

[Table 27 Best\_NODC7 entry issues example 27](#_Toc403656376)

[Table 28 Incorrect taxonomic system code example 28](#_Toc403656377)

# Abstract

BioChem is a national DFO database that holds biological and chemical oceanographic data from departmental research initiatives as well as from areas of Canadian interest. These data are available internationally via the internet to two hundred and fifty-two registered users. A lack of resources has resulted in limitations to the tools available to load data to the archive. Concerns over the content of the database have been expressed by users and data providers; these concerns led to the audit exercise that is reported here. While examples of violations of the relational database design, improvements needed to the database and associated tools, and a lack of coherent standard practices have been revealed, a number of issues affecting data quality have been discovered. These include the following:

* Duplicated data, including cases where the same data type from the same sample are loaded with different values;
* Same data stored in differing units of measure, without proper description of those units;
* Instances where data collection date is uncertain;
* Multiple cases where mission start and end dates are incorrect;
* Instances where unit conversion was done incorrectly, or the method is not known;
* Instances where nulls were set to 0;
* Typographical and duplicate entries in look up code tables;
* Inclusion of data for which the database structure is inappropriate, resulting in columns intended to house one type of information actually containing something very different (ex: Pigment Colour Index housed in the Plankton ‘counts’ column);
* Data associated with the incorrect metadata, i.e. data from one mission loaded with another one from a different year;
* Inclusion of known erroneous data without setting the QC code to reflect that;
* Data erroneously changed from original value before load (e.g. 0s stripped out of DOXY, for example 207.5 became 27.5);
* Missing data, it appears that at times not all provided data was loaded;
* Questionable sample IDs associated with data (e.g. single digits, negative values, etc.);
* Mysterious cases of repeated, erroneous values (e.g. multiple DOXY samples with identical values, that do not match with provided preload data);
* Cases of corrections being applied to the original data after load but these corrections are not reflected in BioChem;
* Cases of data in BioChem which are not in the provided preload files (neither the sample ID nor the data values are in the file);
* Cases of questionable locations.

The issues identified will require a major investment of time and effort to correct. It is important to note that a number of the issues uncovered were as a result of ‘eyeballing’ the data to attempt to address a separate issue, the nature of these issues means that this is the only way to find these.

# Background

BioChem is a relational Oracle database designed to house various biological and discrete oceanographic data collected during a mission. These data are available via the internet to registered users worldwide. For additional background on the types of data, missions, and additional technical details on the database structure refer to the technical report available at <http://www.dfo-mpo.gc.ca/Library/351319.pdf>

Data loaded into BioChem have primarily been loaded by personnel in either Maritimes or Quebec regions. Given a lack of resources after the initial development of the database a proper load procedure for raw data was never developed. Each of the two regions developed their own methods of load to the edit accounts, with Quebec also developing a QC procedure applied prior to the load for discrete data. Within Maritimes the data manager developed suites of MS Access databases and queries in order to massage data provided and ultimately load into the BioChem edit structure. With the retirement of the previous data manager a new team has been looking at this process with the goal of improving and expediting the provision of data to the archive.

This is the second draft of an audit of the content contained within the BioChem database. The need for an audit of the content arose due to concerns expressed by various end users regarding the data which they received from the database, either through queries submitted via the BioChem query application, or through more direct methods querying the underlying database structure (ex. SQL queries).

The ultimate success of a data archive relies on the usefulness of those data to the community of users, not on the total number of records housed in the archive. Users must have some level of confidence that data provided by the archive meet an understood level of quality, are properly described, and are fit for their specific use.

While this content audit has been a focus in the Maritimes region for the past number of months, other priorities and a lack of resources has meant that this is not yet considered complete but should be considered a work in progress. While issues have been discovered and recommendations put forth it will be left to the BioChem Steering Committee to reach agreement on the resolution of these issues and the best way forward. This document may be used to assist in the development of best practices agreed to amongst the regions as well as to inform technical experts on required improvements to the existing edit application and database backend.

It should be noted that the vast majority of the issues with the content involve Maritimes data. While improvements to the archive and associated applications would assist in avoiding some of these, poor data management practices have been the cause of many of these issues.

# Standard Practices

There are a number of fields in the BioChem archive for which there are no clear cut rules for the content. For example at the BCMissions level there is Institute, this field has been populated in a variety of ways. Institute acronym and organizational acronyms are two methods, given the fluidic nature of organization in the Department it seems more appropriate to use an institute identifier and make certain that the mission leader’s name is correct.

More problematic is the population of the collector’s event, deployment, station and sample identifiers. There are entries which ‘appear’ to be what was provided originally, however the majority of the entries appear to be a concatenation of various bits of data, for example the original mission number followed by the original event id, at times the event ID is left padded with zeroes following the mission number at other times it isn’t. Some of these may contain underscores, at times double underscores, to break apart the concatenated bits of data.

This is problematic for end users as there is no straight forward method to link back to their originally provided data to assist them in verifying the quality of those data or to enrich their analysis with data from the same event/sample which may not be loaded into BioChem. It is also the case that the same event has multiple records at the event level, with no easy method to QA the contents between one version of the event record and the other it is likely that the same event is loaded with different entries for the collector’s event and/or sample ids.

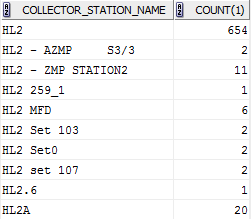
This lack of standard is problematic for end users, for example the variations in the station identifier make analysis of these data problematic. Below is an example of some of the variations possibly indicating Halifax Line Station 2, there are other instances in the archive which likely also indicate this as being the station, Hfx Stn 2, H2, etc. This field does not lend itself to being constrained to a look up table due to the nature of the data. Standards need to be applied and followed in the provided data and the load process in order to facilitate use of the data. The same can be said for the sample id.

Table Collector station name example

# The Data

## Raw Data Archival Issues

During this audit there have been a number of times where referencing of the original data provided was attempted but not possible. There have been a number of network locations where data either loaded, or intended for load, have been located. However it appears that there are a large number of files missing. Other issues with the current raw file archival are; multiple versions of the same file, cases where the files were updated after load into BioChem and these changes were not then applied to the previously loaded data, and multiple locations containing data and metadata.

Below is one example of the difficulty in attempting to address possible issues with the flat file archive in the state it is in, as well as the practices associated with the management of these data.

During analysis of dissolved oxygen issues ten cases were discovered during a comparison of one mission between the original data and the loaded sample IDs and data values where neither the samples nor the values were in the original file (HUD2000050). Further investigation of these particular samples show a number of various discrete data loaded. Given that these were not in the original file it was considered possible that they belong with a different mission. A query of the database for the sample IDs immediately prior to and after the sequence are associated with Halifax Station 2 fixed station missions for 2000 and 2001. A check of the original file shows that these samples (213371 – 213380) were actually collected on October 25, 2000 at Halifax Station 2. However it is not clear where the files containing these oxygen values are now, the conformation of the sample IDs was based on Chlorophyll data.

Nowhere in the flat file archive for HUD2000050 is there any indication of other data collected after October 16.

In checking the dates of HUD2000050 in BioChem the last date is October 25, however the last data collected on that mission prior to that was from October 16. Checking the Cruise Project Database shows us that there was a second leg of HUD2000050 with Mitchell as the PI, the first leg having ended on Oct 16. Double checking the bridge logs shows that the second leg did occupy Station 2 on October 25th.

## BCMissions

The BCMissions table houses information on the mission from which the data originated. Each mission may have multiple events that took place.

Current record count is 2,847. Note the current estimate of unique missions based upon descriptor is 2,109; this number is inflated as there are typos in the descriptor field.

### Protocol, In the Wrong Place?

Currently protocol is housed at the mission level. Given that opportunistic sampling occurs during many missions, for example AZOMP missions pick up stations for AZMP, it is important to have something regarding this at the sample level, this allows users to quickly identify those samples of interest to them. For the time being Maritimes has begun to include an indicator in the Collector column for this info, however this is not ideal. Consideration should be given of new columns at the BCDiscreteHedrs and BCPlanktnHedrs levels with an associated lookup table.

### Incorrect Data Center Codes

The data center code is used to identify the region/institute of the custodians, collectors, and managers of the data for the mission. This allows users to know where to go initially with questions, query data for specific institutes, and to assist data managers in identifying their data. This code is also used in order to hold back data from the BioChem query application, only specific codes are available to the query application.

It appears that during a number of loads of data on behalf of other regions by Maritimes the data center code was mistakenly set to BIO.

This happened forty four times for NAFC data, nine times for IOS data, and once an IOS mission was given the NL data center code. With regards to IOS data the appropriate data center code was assigned only twice out of twelve records, both of those having been loaded by an IOS resource. BIO data were assigned a data center code for NAFC five times.

The BioChem edit application does not allow for an update to this code.

Further to this BioChem also contains SAHFOS data which DFO is not licensed to distribute. The method for keeping these data from public access is the assignment of the code ‘21’ there are currently 44 mission records associated with this code. The name for these records is only populated three times, each time it has a prefix of ‘cpr’. The descriptor appears to be year followed by 30002. Analysis was performed based on these two business rules and it appears that the SAHFOS data has not been given an incorrect data center code. Meaning that there are only forty four missions which fall under this agreement, and that they are not publically available.

Corrections for this issue had to be performed by running SQL scripts against the BioChem production table, BCMissions. This is not a recommended practice. These corrections were applied in August of 2014. Scripts and note of the outputs have been saved, either in mission specific documentation or in the general BCMissions data fix archive at BIO depending on the nature of the discovery of the issue.

It is recommended that any replacement of the current edit application address this issue.

### Duplicated Mission Records

Ideally, in a relational database, this table would house one record for each mission. This facilitates maintenance of the quality of the data provided for the mission, for example if an error is identified and corrected in the mission information it need only be corrected once. However, due to the nature of the BioChem Edit/Load application this relational rule has been violated consistently in the archive. The application treats data collected during a single mission as two different functional areas, Plankton and Discrete. The body of data for each of these areas are treated in essence as different missions, regardless of the fact that they were collected during the same mission. Thus if both plankton and discrete data were collected during the same mission, this would result in at least two mission records from the same mission. Unfortunately the allowance of this violation of a basic tenant for relational database methodology has led to many other issues and duplications in the archive.

Further to this issue there is no constraint in place either through the edit app or the database structure which keeps a data manager from loading the same mission multiple times for the same functional area, each individual mission record may have differences in the content, it should be noted that these duplicates are not necessarily indicative of duplicated data, but do result in duplicated event records. At best this erodes the confidence of the end user and adds greatly to the difficulty in maintaining the quality of the data. At worse this has allowed data collected during a single mission to be loaded more than once, resulting in duplicates of the same data type and value for the same sample. Duplication of the same data results in skewing any analyses.

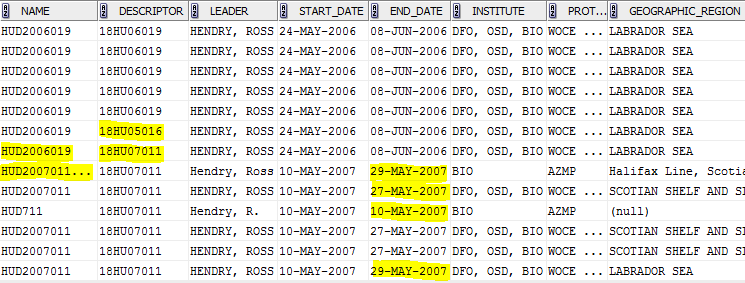


Table Duplicate missions example

One of the issues in the above figure is for mission name HUD2006019. This particular mission was reviewed in August, 2014. The review included careful examination of the original data provided and the data as loaded into BioChem. This showed that the two trips with the descriptors of 18HU05016 and 18HU07011 actually were from those missions in 2005 and 2007, in this case the mission information was incorrect rather than the descriptor. An examination of the data loaded as names of HUD2005016 and HUD2007011 confirmed that the data loaded as above were not duplicates of data already associated with these. The datasets were downloaded and the mission information corrected and reloaded, the full details were documented.

Additional analysis also showed that one of the entries above for HUD2006019 with 18HU06019 was an exact duplicate of data already associated with 18HU05016 yet with all of the mission information being for 18HU06019. That duplicate (mission\_seq 20000000003860) was deleted from the archive to remove the duplicated data. The mission records after this cleanup for this particular mission is shown below. Also see Data Loaded to Wrong Mission.

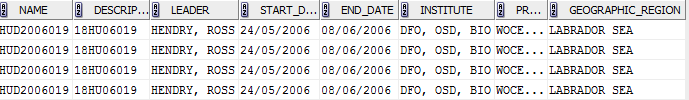


Table HUD2006019 missions corrected

### Trailing White Spaces in Name and Descriptor Fields

There were a number of cases where trailing white spaces were found in either the mission name or descriptor (377 records). Oracle views a white space as a valid character and these resulted in some records not being returned when a query didn’t take this possibility into account. For example querying for HUD2007011 would not return the mission record shown in Table 2 where the white spaces are represented by ‘…’.

A script was applied which trimmed off these problematic white spaces. Additional analysis should be performed to identify this possibility in other fields which are used for searches.

### Incorrect Start and End Dates

It appears that a number of times when loading data the mission’s start and end dates were not known. During the load the start date was assigned, perhaps based on the first event, and at this time the end date for the mission was simply identical to the start time. This was done in 1,097 cases. While we might assume that in some cases this is valid a great many of these have data associated with them which fall outside of the recorded mission dates. This makes analysis and quality assurance difficult.

Experience has taught us that, at times, data were incorrectly loaded to a mission, for example HUD2006019 (see Table 2) actually had data loaded to it which belonged with a mission in 2007 and another mission in 2005.

Further analysis is required in order to correct the date issues at the mission level, or to correct the mission records where the data actually belong to another mission.

Attempts will be made to identify the correct start and end dates of the missions where possible. A great deal of the data loaded in Maritimes for the Newfoundland Labrador region appears to have identical start and end dates. The region where the data originated will be queried to see if appropriate dates can be assigned.

Where it is not possible to find the original mission information, a review will take place of the associated data and decisions made on an individual basis as to whether or not use of the first and last sample dates seems prudent.

There were a number of cases where the month and day of the mission dates were flipped, for example March 9 should have been Sept 3. There were also cases where the start and end dates were reversed. These were addressed on a case by case basis. It should be noted that the identification of these cases relied on various analyses, feedback from other regions, tests against the sample or event dates, and for Maritimes data tests run against ODIS’ Cruise Project Database assisted with identification of these issues.

### Data Loaded to Wrong Mission

While attempting to address another issue with a mission (HUD2006019) it was discovered that discrete data associated with this mission were instead collected during HUD2007011. While correcting this issue it was then discovered that discrete data under HUD2006019 was actually from mission HUD2005016. Currently it is not known how often this may have occurred. It is possible that these issues arose due to re-use of procedures containing hard coded values which were not appropriately revised.

## BCComments

This table is used to house additional comments for either the BCMissions or BCEvents tables.

Current record count is 13,945.

It was intended to be used in order to provide additional information on such things as protocols, original data sources, issues with instrumentation, etc. which extended beyond the 2000 character limit for the data manager or collector’s comments fields.

There appears to be confusion as to what this table should house. At times it includes what may be an indicator of a fixed station id, which may or may not also be included in the BCEvents table. And at times these station indicators are in conflict between the two tables.

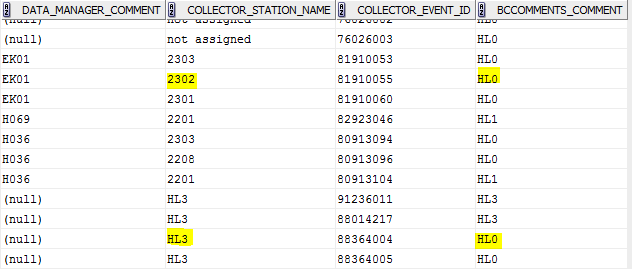


Table BCComments and station info

There are also 6,004 cases where a record exists however the only comment is ‘N’ or ‘0’. If this is meant as an indicator that no additional comment exists these records should be removed in order to avoid confusion.

## BCEvents

The BCEvents table houses information about an event. An event is considered as something happened here. For example a vessel stops on station and deploys an instrument. Each event may have multiple headers associated with it in either of the two functional areas, Plankton and Discrete. An event could be considered as bound by start and end times and minimum and maximum positions.

Current record count is 151,003.

### Duplicate Events

There are 9,441 events which appear to be duplicated in the archive. The script tests based on start date and time and the collector’s event and station information. These may be the result of the loading of the discrete and plankton data separately, or related to other downstream duplicates noted in this document.

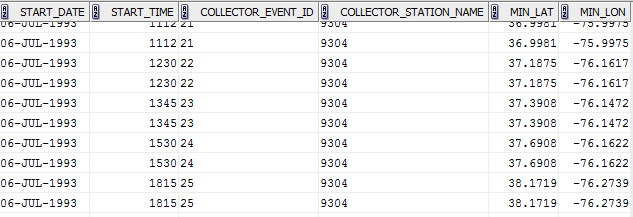


Table Duplicate events example

### Dates Outside of Mission Dates

There are currently 25,001 records where the start date of the event does not fall between the start and end dates of the mission with which the event is associated. This brings to question which of the data are in error. When such a simple thing as ‘when’ is questionable it reflects poorly on other aspects of the archive.

This issue could be resolved for the long term by the implementation of a check against such occurrences before the data are loaded to archive.

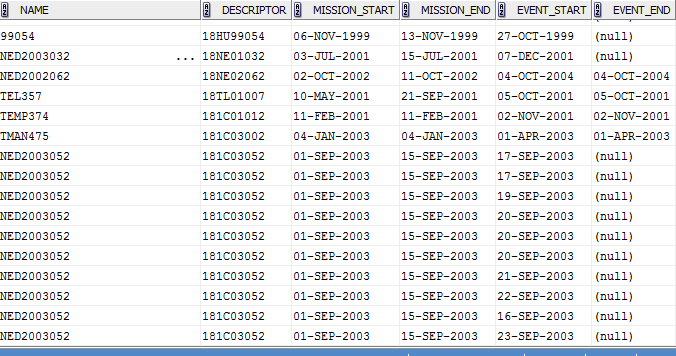


Table Event outside of mission dates example

## BCDiscreteHedrs

This table houses information on the discrete sampling activities associated with an event. For example samples from bottles are a discrete sampling activity.

Current record count is 663,711.

### Duplicate Discrete Headers

There are 24,496 duplicated records in this table. The script tested for duplicates based on gear, start date, time, depth, position, and the collector sample id. Again these may be the result of multiple loads with different data types, or related to duplicated discrete data. Additional analysis is required.

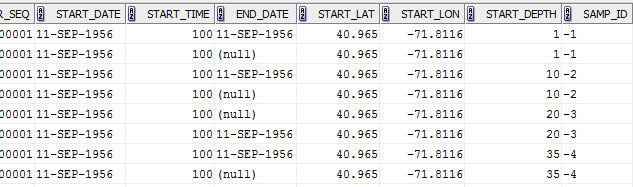


Table Duplicate discrete header examples

### Dates Outside of Event Dates

As is the case with the BCEvents table there are currently no edit checks implemented in the edit application that ensures that a sampling event, or header, takes place within the temporal boundings of the event with which the activity is associated. There are currently 19,990 records where the discrete header’s start date is outside of the start and end dates of the event.

It is apparent that in some of these cases this is due to a date format issue with the data as provided to the system. For example the sampling appears to have taken place on July 11th, and the event took place on November 7th.

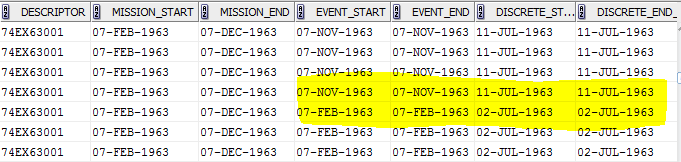


Table Discrete header dates outside of event dates example

### When Depth is not a Depth

An earlier data rescue exercise resulted in the inclusion of analysis completed on sediment samples into the BioChem archive. Looking at the depths associated with these data it appears that at times the depth fields were used to indicate the depth into the core sample, at other times the samples may have been collected by a grab and the sounding depth appears to have been used, and at yet other times the depth is set to 0. The use of 0 seems to indicate that the sample was taken from the top of a core, so in this case rather than it being the surface of the water it represents the bottom of the ocean. Data associated with grabs appear to either be the sounding depth, or somewhere else, either in the water column or perhaps depth into the bottom. It is not clear at this time.

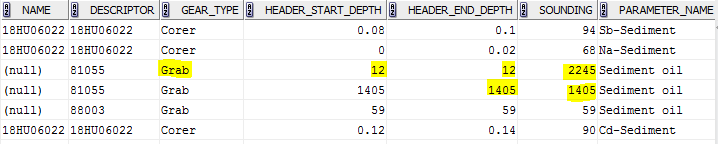


Table Sediment data and depths example

This is a result of shoe horning data into an archive which it was not designed to house. The current model is focussed on the water column.

This issue is also present in the Ice sampling data rescued and loaded into the database, see When Gear is not a Gear.

Further investigation is required to properly document these cases.

### When Gear is not a Gear

A data rescue project focussed on ice studies resulted in these data being loaded into BioChem. There are a number of issues with this. The field for sampling gear was used to indicate ‘Ice’, with the variations of it being an ice sampler, ice chips, the top of the floe, the middle of the flow, or the bottom of the flow, or no description at all of the meaning for ‘Ice’. The depth fields were set to 0.

To end users looking for physical measurements at the surface this means that it is possible that the result set could include those from ice without them being aware of it. Again, this is a case where the database was used to house data for which it was not built. The current model is focussed on the water column.

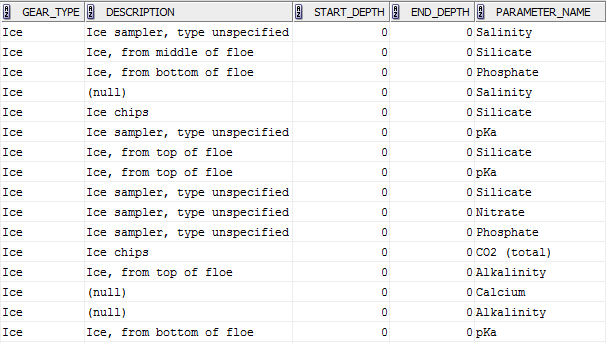


Table Ice gears, depths, and parameter example

As of August 2014 thought is being given to clone the BioChem applications and database to allow ongoing archival and access to these data, while making it obvious to end users what it is that they are getting. This would allow for these data to be removed from BioChem. IM&TS is recommending submitting a request for the development of a new database and applications.

Further analysis is required to resolve this issue.

### Questionable Positions

A quick check against the discrete data shows that there are 109,718 cases where minimum and maximum event latitude is recorded but the latitude associated with the data fall outside of that range. And 111,068 records where the same issue appears for the longitude. This might be an indication of sample data loaded to the wrong mission, or possible issue with the locational information supplied at the event level.

## BCDiscreteDtails

Houses information for the various parameters collected during a discrete sampling activity.

Current record count is 3,582,110.

### Duplicated Discrete Data

There are 25,696 cases in the archive where the same data was loaded at least twice. The script tested for cases where the date, time, start positions and depth, collector’s sample id, and data type are identical. Note that this will not identify such cases as where times were revised prior to a second load.

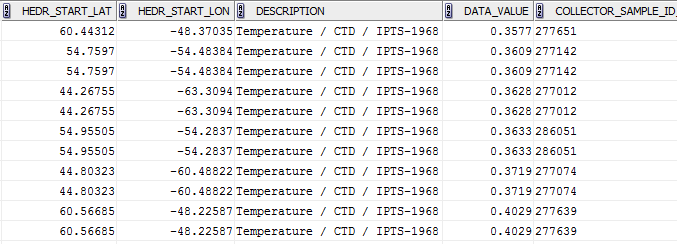


Table Duplicated discrete data example

There are also cases where the same data was loaded more than once, but the value was converted during one of the duplicate loads. In the example below it appears that the dissolved oxygen was loaded once with a conversion applied, and again without.

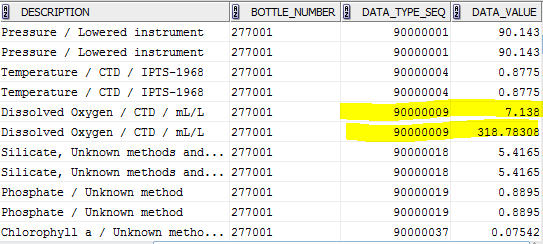


Table Discrete data converted and not converted duplicate example

### Dissolved Oxygen Conversion Issues

There are a number of data types in the archive which are related to dissolved oxygen.

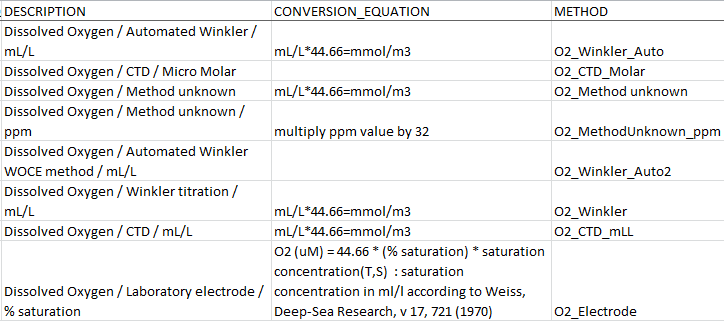


Table Dissolved oxygen conversions

Conversions have been applied prior to load to the archive for many of these. IML’s process has been to convert using mL/L\*44.66, while in Maritimes the conversion applied was mL/L\*44.6. The ICES site states that the standard for this conversion is mL/L\*44.661. It has also been discovered that some of the dissolved oxygen values converted prior to provision of the data in Maritimes made use of yet another conversion of mL/L\*44.5. This will complicate the correction of these values.

The unit associated with these converted CTD values is described in the archive as mL/L, however the query application states that the unit is mmol/m3. There have also been cases where the data were loaded without any conversion applied. Given the differences in conversion practices use of these data should not be encouraged until the issue has been resolved.

Best practice would be to load data as provided, ensure that it is properly described in the archive, and allow the users to apply conversions to the data as they see fit for their purposes. This allows for greater ease of managing the data, for example the value in the archive is the same as the value supplied by the provider.

In an effort to try to get a handle on this issue two views were built with the assumption that a value of 12 might be considered a reasonable delimiter on whether data were converted to mmol or not. In addition a review was conducted on various versions of the MS Access database, the associated queries, and the original data used over the past 13 years in Maritimes in an attempt to confirm that a conversion was applied. In all cases where this intermediate step was available it appears a conversion was applied, however in many cases either the original data or the associated Access database could not be found. Much more analysis would be required in order to ascertain exactly what is happening before applying any corrections to the data. For example it appears that no conversion has been applied to the % saturation values but this needs to be confirmed.

Using these views a query gives a quick snapshot of possible cases where no conversion was applied to these data at all.

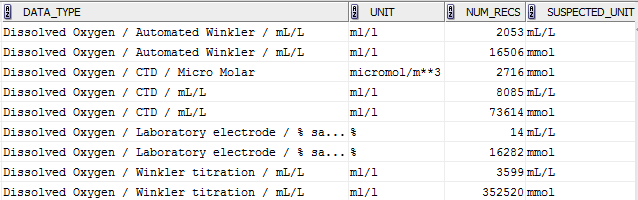


Table Dissolved oxygen conversion issues

In the fall of 2014 the Atlantic Zone Monitoring Program in Maritimes provided funds for a short term contract to look into the oxygen issues for Maritimes’ AZMP, AZOMP, and the related fixed station and Scotian Shelf surveys. The results of that analysis can be read in the document DOXY\_BioChem produced by Gordana Lazin.

### Data of Unknown Origin

While attempting to resolve questionable start and end dates for missions in conjunction with ISDM it appeared that some of the data may actually be associated with the wrong mission (ex. 181C05626 ) based upon PI names and dates it appears possible in this particular example that these samples were from two other missions (181C05634 and 181C05654). However the original data cannot be located at this time in order to confirm that this is the case.

### Nulls Set to Zero

There are cases when there was no value for a parameter in the spreadsheet provided and the null value was set to 0 before loading to the archive. As it is conceivable that the value of 0 is valid this issue needs to be addressed somehow and the incorrect values removed from the database. This would mean identifying cases of 0 in the archive, reviewing the data as originally provided and removing those records found to be in issue.

### Negative Values Set to Zero

In Maritimes region it became apparent that there has been a practice of revising the value of data provided at times. In this particular case for Phaeophytin, where the value was a negative number the value is reset to 0. This was perhaps done due to the range check within the archive which does not allow for a negative value for this parameter. A better practice would be to revise the minimum range and load the data as provided, setting the quality assurance flag to reflect that the value is questionable and possibly erroneous and to include the detection limit for the instrument.

Further analysis needs to be performed to ascertain if this practice has also been applied to other parameters.

### Zeroes Stripped Out of Data

While looking at oxygen conversion issues a comparison was made between the values loaded into BioChem and the values originally provided. The mission was HUD99054, this exercise resulted in the discovery of a few issues, one of which is that it appears that prior to load to BioChem all zeroes were removed from the file.

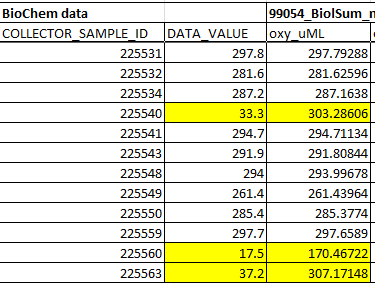


Table Stripped out zeroes

### Previous Data Values Assigned to Nulls

Again while comparing data for mission HUD99054 it was noted that at times when there was no data for oxygen for the sample, a value from previously in the file was somehow assigned to that sample. Note that without a visual comparison of every data file to the loaded data it is impossible to identify this issue. In the data shown below each of the different coloured rows represent cases where that value was ‘recycled’ later in samples where no oxygen values exist.

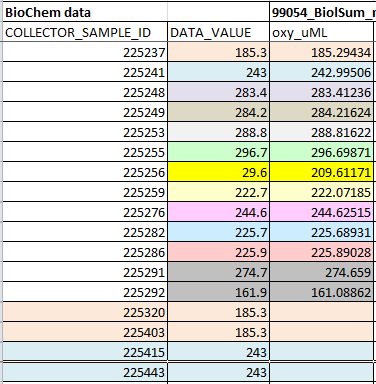


Table Example of ‘recycled’ data values

### Not All Provided Data Loaded

Again while comparing data for mission HUD99054 it was noted that at times data provided for a sample somehow were not loaded into BioChem. The nulls shown below indicate no oxygen data for that sample is in the database. Note that without a visual comparison of every data file and the loaded data it is impossible to identify this issue. This issue has also been identified a number of times through other work, at times it is the entire dataset (ex all HPLC), at other times only some of the samples are missing.

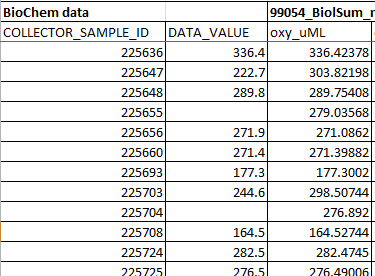


Table Examples of missing sample data

### Repeated Values for Different Samples

In a review of oxygen data for HUD2001022 it was noted that somehow identical values were loaded for a number of different samples, this did not happen for all of the samples. In the figure below the first column is the sample ID, the second the value in BioChem, and the third contains the value in the original data.

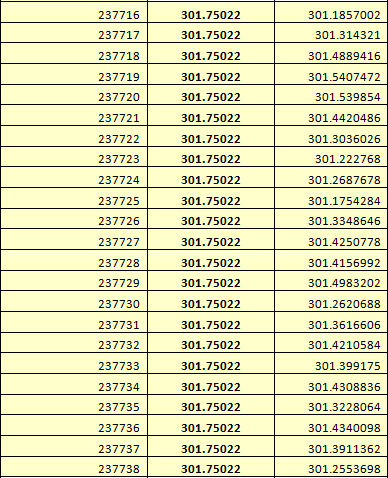


Table Repeated values

## BCPlanktnHedrs

Houses information regarding plankton sampling activities associated with an event.

Current record count is 105,243

### Duplicate Plankton Headers

There are currently 184 records duplicated in this table. The script tested for duplicates based on gear, start date, time, depth, position, and the collector sample id. Further analysis is required in order to ascertain whether or not these are true duplicates or the result of multiple loads of the same mission with different data types.

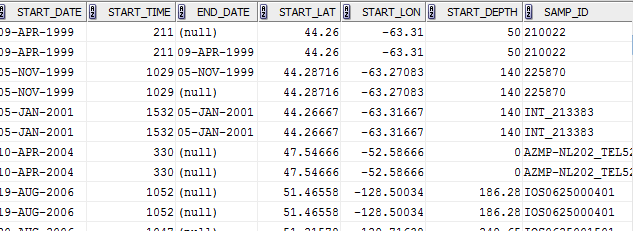


Table Duplicate plankton header example

### Plankton Dates Outside of Event Dates

As is the case with the discrete header data this table also contains dates outside of the event dates. There are 1,328 records where the plankton header start date does not fall within the start and end date of the event.

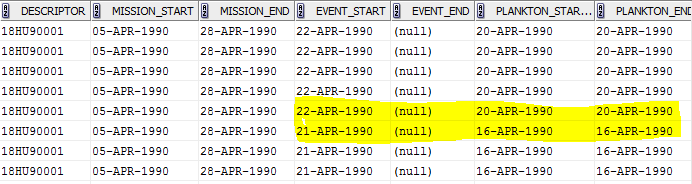


Table Plankton header date outside of event dates example

### Nominal Positions

While reviewing NL’s plankton data it was noticed that a great many of the samples for Station 27 appeared to have identical geographic information. It appears that this data may have been ‘fudged’. Perhaps it was not originally supplied, or was absent from the provided plankton information and a default was assigned instead of tracking down the actual location.

Out of 11,594 plankton sample records for NL at station 27, 10,016 have a latitude of 47.54666 and a longitude of -52.58666.

This is the result of making use of standard positions for fixed station sampling, rather than the true position of the sampling. The reason for this practices appears to be simply that this is the way it’s always been done. Given the improved GPS technology now available it should be considered whether or not this practice should continue. It is not readily apparent to end users that a position in the database is nominal versus actual.

### Questionable Positions

A quick check against the plankton data shows that there are 124,368 cases where minimum and maximum event latitude is recorded but the latitude associated with the data fall outside of that range. And 106,837 records where the same issue appears for the longitude. This might be an indication of data loaded to the wrong mission, or possible issue with the locational information supplied at either the event or sample level.

## BCPlanktnGenerals

This table houses data based on the results of the analysis of the sample, such as species identifier and counts.

Current record count is 1,583,758.

### Duplicate Plankton Data

There are currently 263 cases of duplicated records in this table. The script tests for those cases where the start date, time, position, depth and the sample id, taxonomic, life history, trophic codes, and the sieve size, split fraction and sex are all identical. Due to the complexity of getting at these data a number of scripts were required in order to pull the duplicates.

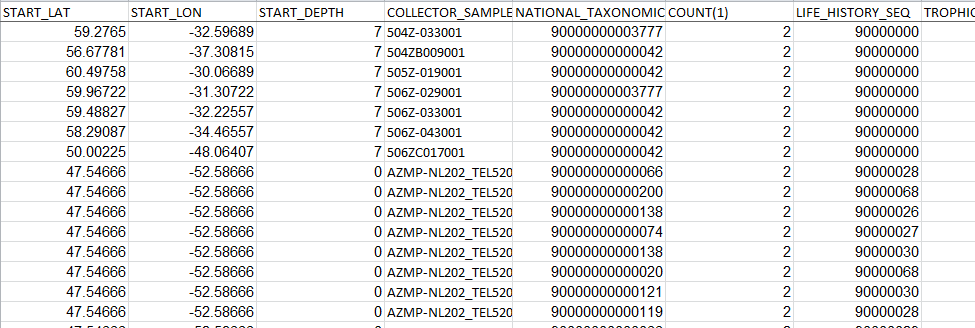


Table Duplicate plankton example

### When Count is Not a Count

Included in the BioChem database is a copy of Continuous Plankton Recorder (CPR) data provided under a licensed agreement by the Sir Alistair Hardy Foundation of Ocean Sciences (SAHFOS). These data are provided with either a true count of animals, or a representation of an estimate, for example when the count is 35,000 no one actually counted that many animals. However at other times the number may reflect a true count.

Further to this complication these data also include Pigment Colour Index (PCI) codes for phytoplankton. Again this value is loaded into the column intended to house the count of individual animals. The values used for this are 0, 1, 2, or 6.5. All of which may also indicate a count. In addition to this and further compounding the use of these data, none of the records with a value of 0 have been loaded only the station level data. The researcher must create the 0 records themselves based upon an absence of any other PCI for a station. *It should be noted that for these issues there currently exist seven records of data of Maritimes’ data, not SAHFOS. Additional investigation is required in order to ascertain if these are true PCI values or perhaps counts of cells.*

Any end user of these data needs to be aware of these and other nuances before making use of the data. Serious thought should be given as to whether or not the BioChem model is the appropriate archival and access tool for the DFO user community.

# The Lookup Tables

## Lack of Lookup Tables

Lookup tables are typically used in a relational database in order to assist data managers in maintaining the quality of the data. For example this insures that typos do not end up in the archive. These also facilitate querying of the data and assist the user in getting everything out of the database which they expect.

There are a number of columns which would benefit from the addition of a look up table.

For example at the BCMissions level perhaps the addition of a lookup table to assist in maintaining the platform name in a consistent manner would be beneficial.

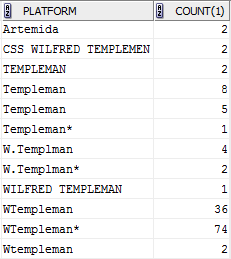


Table Platform variation example

The collector station id is another column which, if the contents were properly managed, the users would find very helpful. Implementing a look up table for these data at present is not possible as there are cases where a station id is not standard. It should be considered whether or not the inclusion of a new column meant to house only standard station ids should be included and referenced to a lookup table.

## Lookup Table Housekeeping

At times entries exist which are simply confusing, a review and better descriptions of the meanings of these needs to be performed. Below is the content for the BCSexes table.

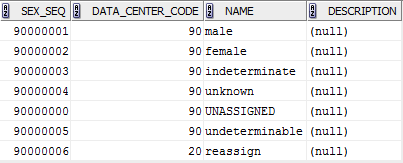


Table Confusing lookup table example

Within a number of the lookup tables there are entries which appear to have been added accidentally, these may have the content of ‘Database artifact’, ‘Delete’, ‘Disregard’, or ‘Reassign’.

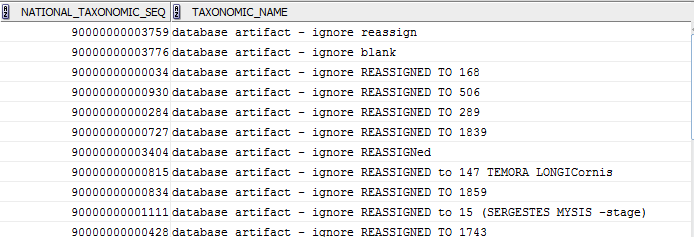


Table Invalid lookup table content

Extraneous entries in the lookup tables increase the chances of data being accidentally loaded with an associated code which is actually invalid. In order to assist in maintaining the quality of the data these should be removed. For example, there are fifty plankton records in the database which point to one of the above invalid taxonomic names.

## BCNatnlTaxonCodes

This lookup table contains information on the species.

Current record count is 5,487.

### Content with No Associated Data

There are currently 2,846 records in this lookup table which have never been used in the database. These additional extraneous records add to the complication of managing this table and the associated data. Some of these unused records are actually duplicates of existing records which are associated with data.

### Authorities Which Aren’t

A review of the authorities in this table should be performed, there are apparently cases where a species was added and coded, but the authority provided was not actually an authority. See calanoida Mauchline 1988.

### Duplicated Records

There is a serious issue with duplicated entries in any lookup table within a relational database. As the table is meant to facilitate maintaining the integrity of the data this practice is an issue. At times these duplicates may not be exact, for example there may be some small variation or typo in the spelling, or at times a common name was used where another record exists for the scientific name.







Table Duplicate species examples

The below example shows three records for Euphausiacea, the only difference in these records are the data manager’s comment and the assigned taxonomic code.

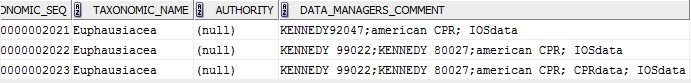


Table Using lookup table to track source example

In the next case while the most meaningful fields have been duplicated there is a difference in the Best\_NODC7 entries.



Table Best\_NODC7 entry issues example

All of the above issues cause problems for end users, for example if someone were looking for data on Cirripedia they may not realize they need to include the additional taxonomic code for ‘barnacle’ in the query. For the data providers and data processers it requires a decision be made as to which of the multiple codes should be assigned to the data. In Maritimes it is the code with the most data associated with it. For each of the examples provided above data exists in the archive.

Given the variations in spellings of both the taxonomic name and authority column, as well as format variations in the authorities it is difficult to get a quick handle on how many duplicates actually exist in this table.

### Possible Miss-mappings to AphiaId and TSN

A review needs to be done in order to identify and resolve those cases where the TSN or AphiaId are incorrect.

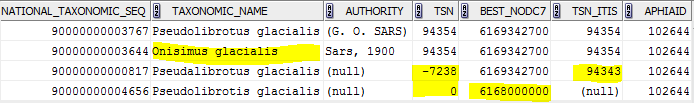


Table Incorrect taxonomic system code example

# The Corrections

During the past year a number of issues have been identified and resolved on a case by case basis.

Where ever possible the original data and sources outside of BioChem were used in the analysis of the issues and to assist in the identification of the appropriate correction. The original data provided and used to load to BioChem were not always found. In these cases the issues were dealt with on an individual basis. The external sources used were typically the Master Mission Table at ISDM, the AZMP website, and ODIS’s Cruise Project Database.

Generally the corrections took one of two forms. Either a script was written and applied to the production tables, or the edit application was used.

While it is preferable and recommended to use the edit application where possible this was not always the case, either due to the sheer number of and/or nature of the corrections, or a lack of time/resources.

All deletions of missions, for example in cases where the mission and the data were duplicated, were always performed via the edit application in order to be certain that all records were removed.

All corrections have been documented and archived within ODIS. The documentation and supporting files can be found in the BIODataSvcPrv share on the ENT domain, located in 600\_Production\BIOCHEM\BioChem\_Data\_Fixes

The type of documentation may vary depending upon the nature of the discovery of the issue. The documentation generally follows one of two paths 1)individual documents and supporting files for either a mission name (ex HUD2009011) or a descriptor (ex 18HU09011) filed in a folder matching with that name or descriptor, this may include a wide variety of issues and corrections 2)analysis against a larger portion of the database, for example BCMissions, in this case the documentation will be broader, and typically focussed on one issue only, but will also include the results of analysis and associated files and scripts.

Corrections of the same nature, for example incorrect data center code, may be found in either one of these places. This is the result of errors being noted while analysis of another nature was being performed; if time and resources permitted the correction was applied during the other analysis. If not then a note was made in order to allow for follow up at a later date. The analysis required in order to be confident of the appropriate correction can be very time consuming, and therefore disruptive to the analysis being undertaken at the time.