



# The BCO-DMO Metadata Database Schema

Notes:

1. All tables have a created\_date field.
2. The platform table replaces the cruiseid table to be more general. It could also have been called the deployment table.
3. I decided not to change name specific id's, like people\_id to be just id even though it works well elsewhere. The problem that I see is that it will make reading linking tables harder to understand.
4. The location table contains one or more positions/times for this dataset, getting the data from the event log perhaps or from the meteorological data sets.
5. Do we need the tables for instrument and dataset\_instrument? - Yes.
6. Do we need to store max and min values of positions, and other parameters? STILL TO BE DECIDED
7. There will be more than one location table entries for a dataset. Hence the need for a linking table, dataset\_platform table.
8. June 21, 2007. Replaced depth with depth\_w in location table.
9. June 26, 2007. The location table is now connected to the Platform table, representing a summary of the “best” navigation for the cruise. We need to encode information about the starting and stopping of measurements and that is done using the new Start\_stop table, mainly of dates/times. This table does include fields for latitude and longitude, but if these are not present, the system should look up the position information in the location table, dead reckoning between positions if necessary, assuming a straight line between fixes. To reconstruct the cruise track, it will be necessary to sort the retrieved information from the location table by time (or maybe location\_id if the data are added in order). It would be best if the times were stored in the location table as GMT, with timezone information included with each record. This information may be used to convert input data provided in local time to be stored in GMT. It was noted that we do not have a “Scientific\_personnel” table to record who participated during the cruise. This could be added, but at this time it is felt that names connected to each of the data sets, via the Dataset\_platform\_people table, is sufficient. Another discussion to have is whether tow, cast, and/or station number information should also be saved within the start\_stop table. These are currently included, but ultimately the overhead in storing and maintaining this information may be too high. Note that the Start\_stop table contains the platform\_id, which is redundant information, given that it could be obtained via the dataset\_platform\_id value, but platform\_id is provided as a way to simplify subsequent retrieval.
10. September 13, 2007. Added dataset\_platform\_url to the dataset\_platform table. Add the dataset\_status and dataset\_platform\_status tables. The “status” field name in both tables should probably be an enumerated list while the comment list will enable more free form information.
11. October 10, 2007. Add Project and Program tables. Replace project with project\_id in the Dataset table. Clarify that the status information in the Dataset\_status table is an ENUM list and therefore a controlled vocabulary. Add acronym field names to the affiliation, funding and instrument tables. Possible ENUM values for the title in the Dataset\_platform\_people table are originator, contact, analyst, technician, curator, manager, student, principal investigator, scientific investigator, and publisher.

12. October 11, 2007. Remove dataset\_id from the Location table and change the dataset\_platform\_id to platform\_id in the Location table. In the Platform table, the deployment field is a non duplicates field. Add synonyms to this table. It will be a comma separated list of synonyms for deployment. Add a version field to the Dataest\_platform and Dataset tables. In the Dataset\_status table, add entered\_by\_id and next\_action\_date fields. In this table, the comment field will contain potentially lots of text so it should be able to accommodate many characters. Add program\_id and contact\_id fields to the Project table. Add geolocations to the Program table. Remove the conversion\_utility field from the Parameters table. The geolocations field in the Project and Program tables is a comma separate list of geographical areas (e.g. Pacific Ocean, Georges Bank, etc.) but could also contain Marsden Squares and/or C-squares numbers.

13. January 23, 2008. Remove the latitude and longitude fields from the Start\_stop table. These can be determined by using the start\_date and end\_date values and doing a lookup and possibly interpolation to the data in the Location table. Add entered\_by\_id and next\_action\_date fields to the Dataset\_platform\_status table. Fix spelling of this table name. The status field in this table is an enumerated list, similar to the status field in the Dataset\_status table.

14. January 24, 2008. Remove the dataset\_id from the Modification\_history table as this value may not always be ready. Separate out the table name and start and end local ID value from within the description field from this table and put them in their own columns, table\_changed, start\_local\_id, and end\_local\_id, respectively.

15. April 22, 2008. Remove unnecessary dataset\_id field from the Project table. Add co\_pi2.id to both the Project and Program tables. Fix the spelling of dataset\_parameters\_id in the Dataset\_parameters table. Add related\_projects column to the Project table and related\_programs column to the Program table. Add the new table People\_status, comparable to the Dataset\_status table, but for people. Yet to be done is the implementation of the new columns to store the information about how to plot the different datasets. It should also be noted that while the status field in the three status tables are listed as of type ENUM, this has yet to be implemented in the actual tables. In order to proceed with this, we need to define what values this field can take on.

16. May 6, 2008. Change related\_program and related\_project to affiliated\_programs and affiliated\_projects. Add coordinated\_platforms to Platform table. Fix spelling of end\_local\_ed to end\_local\_id in Modification\_history table. Add first\_name\_synonyms, middle\_name\_synonyms, and last\_name\_synonyms to the People table to handle the situation that people change their names. Fix the headings of each box to be consistent, i.e. ending in "table:". Fix the formatting of the Modification\_history table: the entries needed to be left justified.

17. June 14, 2008. Change the MySQL database engine to InnoDB from MyISAM in order to support foreign key declarations. Add in the foreign key declarations for all foreign keys except for the people table. Add a small\_logo\_url column to the Project and Program tables. Add the Lookup table and replace several ENUM columns with lookup ids into the Lookup table. This applies to the People\_status, Dataset\_platform\_people, Dataset\_platform\_status, Dataset\_status, Dataset\_platform, and Platform tables. Note that until the new lookup values are added the original ENUM fields will be retained in the database. However, once the lookup values are added, the original ENUM field columns will be deleted. Change the entered\_by\_id field in the Modification\_history table to be an INT type and a foreign key as well. The “cast” field name from the Start\_stop table has been removed. It was removed some time earlier from the database, but the schema did not reflect this until now.

18. July 8, 2008. Add a many to many linking table between the Dataset table and the Project table (called Dataset\_project) and between the Project table and the Program table (called Project\_program). Remove the affiliated\_project and affiliated\_program columns once these data have been transformed into the new linking tables. Add a the handle column to the Dataset\_platform table to record the DOI or handle assigned by the data archiving agency.

19. August 25, 2008. Remove Modification\_history table from scheme figure to make room for two new tables. Add the Dataset\_type table and the Dataset\_parameters\_type table. These tables could not have dataset\_type\_lookup and dataset\_parameters\_lookup entries declared as foreign keys since the lookup table must first be recreated as a InnoDB type table. Delete project\_id from the Dataset table and program\_id from the the Project table. Add small\_logo\_url to the Project and Program tables. (This change was done some time ago.) Add affiliation\_id to the Dataset\_platform\_people table, but do not declare as required or as a foreign key until application code is updated. Add conversion\_necessary to the Dataset\_parameters table. It can take the values of either 'yes' or 'no', with a default of 'no'. The lookup table was updated to include the new entries needed by the Dataset\_type table and the Dataset\_parameters\_type table.

20. September 5, 2008. Remove project\_id from the Dataset table and program\_id from the Project table as these were replaced by the Dataset\_project and Project\_program tables. [These columns still exist in the database but will be removed once the new tables are fully implemented in the software.] Add dataset\_id to the Dataset\_type table schema picture and add dataset\_parameters\_type\_id to the Dataset\_parameters\_type table schema picture as these were inadvertently left out of the schema picture when these tables were added August 25, 2008. Fix the line joining the People\_status table to the People table.

21. September 15, 2008. Remove project\_id from the Dataset table and program\_id from the Project table. They are replaced by the Dataset\_project and Project\_program tables, respectively. Rename fill\_value to no\_data\_value in the Parameters table and the Dataset\_parameters table. Rename standard\_name to short\_description in the Parameters table. Remove common\_name and equivalent\_name from the Parameters table.
22. September 18, 2008. Correct the spelling of the “no\_data\_value” entry in the the Dataset\_parameters table on page one.
23. October 16, 2008. Add the many to many table Parameters\_program. Move the contents of Dataset\_platform table's acquisition\_description and processing\_description to the Dataset table and rename the former columns to be unique\_acquisition\_description and unique\_processing\_description. Add description and deployment\_report\_url to the Platform table.
24. November 4, 2008. Apply changes mentioned in item 23 to the live database, in particular create the Parameters\_program table, add acquisition\_description and processing\_description to the Dataset table, rename these columns to unique\_acquisition\_description and unique\_processing\_description in the Dataset\_platform table, add description and deployment\_report\_url to the Platform table, move the coordinated\_platforms column to the correct place in the Platform table, and change the rank column to INTEGER (10) instead of TINIINT in the Dataset\_parameters\_type table.
25. November 14, 2008. Change coordinated\_platforms to coordinated\_deployments in the Platform table.
26. February 11, 2009. The following changes have been made to the test database with the expectation that they will be made to the live database shortly. Add contact\_id and co\_pi2\_id to the Program table. Add rank (as a decimal number) to the parameters table. Brief\_description has been added to the Dataset table. Because of additions to be added to the Dataset\_platform table (described below) it is possible that the validated flag will be removed from the Dataset table. Create a new table called Contact\_status replacing the identical People\_status table but adding keyword\_string to this table. Add current\_state and current\_state\_comment to the Dataset\_platform table.
27. February 12, 2009. Add current\_state and current\_state\_comment to the Dataset table. Make the changes to the live database mentioned in the February 11, 2009 entry. Add geometry\_type\_lookup column to the Platform table.
28. February 26, 2009. The following changes were made to the test database but are anticipated to be made to the live database shortly. Remove the Contact\_status, Dataset\_status, and Dataset\_platform\_status tables. Add the Tracking\_status table. Add people\_status\_lookup to the People table. Update the diagram to show the presence of the geometry\_type\_lookup column in the Platform table.
29. March 3, 2009. Implement the changes mentioned in item #28 and rename tracking\_status\_lookup to dataset\_status\_lookup in the Tracking\_status table.

30. March 11, 2009. Add data\_url to both the Project and Program tables. Replace the missing line between the Dataset and Dataset\_platform tables.

31. April 28, 2009. Add the graphable column to the Parameters table, with values of Y or N. Change the columns current\_state in the Dataset and Dataset\_platform tables to current\_state\_lookup and make them integer values rather than text entries. It was noted that the Dataset\_type table probably should have been called Mapserver\_type table.

32. July 28, 2009. Replaced the Dataset\_platform\_people table with a new table called Person\_role which will serve this function for the Dataset\_platform, Platform, Project and Program tables. That is, it will be used by several tables using the column table\_name to specify which table it applies to and table\_pk\_id for the primary key id from this specified table. Remove references to specific roles from several tables including chief\_scientist\_name\_id and co\_chief\_scientist\_name\_id from the Platform table, and lead\_pi\_id, co\_pi\_id, co\_pi2\_id, and contact\_id from the Project and Program tables.

33. October 26, 2009. These changes were made to the live database August 4, 2009, but due to problems in running this graphics program the diagram was not updated until today. We added two new intersection tables, Project\_funding and Program\_funding, to support the many to many relationship the Dataset table and the Program and Project tables. It was done to support the needs of OCB Project Office.

34. March 17, 2010. The following changes are being made. Add supplied\_name to the Dataset\_instrument table, similar in concept to the supplied\_name in the Dataset\_parameters table, to capture what the contributor calls their instrument. Add version\_date to the Dataset table to insure we have a properly formatted date for the version information. We will initially keep the version column in the Dataset table since it is possible that people will have their own text name (none date entry) for the version information. Add version\_date to the Dataset\_platform table. We will think about a way of filling in this field, if empty, initially using the version/version\_date for Dataset. Add data\_use\_policy\_lookup column to the Project table to keep the id number of the entry in the Lookup table containing the statement of the contributor's use policy. Add minimum\_value and maximum\_value to the Parameters table. It is hoped that this will suffice and we will not need comparable columns in the Dataset\_parameters table. It should be noted that we plan to add “archived” and “restricted” as possible values in the lookup\_table for the current\_state\_lookup value in the Dataset table. We considering how to initialize the current\_state\_lookup and current\_state\_comment columns in the Dataset\_platform table. They have never been used so far (probably because OSPREY does not display these as input options). The program\_name and acronym in the Program table are unique entries. No duplicates are allowed nor can they be NULL. Also, the instrument\_name in the Instrument table should be unique and not NULL. These changes will also be added to the database definition. The affiliated\_projects column was added to the Project table some time ago, but was not reflected in the schema diagram. That has been corrected. An initial study of using triggers suggests that we can add a trigger to the Dataset version\_date column so that it stays current based on changes to the Dataset\_platform version\_date column data. We will investigate this further. Update the program table to reflect the addition of the new column, affiliated\_programs. It too was added some time ago.

35. April 28, 2010. Fix spelling of “geolocation” in both the Project and Program tables in the schema picture. There is no “s”, i.e. the word is not plural. Update the February 12, 2009 entry to reflect the addition of the geometry\_type\_lookup column to the Platform table. Add username to the People table schema diagram. It had been left out. Change the database so that the first, middle and last name synonyms columns display after their corresponding first, middle and last name columns. Move the affiliated\_programs column in the Program table to appear after the geolocation column.

36. July 16, 2010. Change project\_id to program\_id in the schema diagram in the Program table. Add new Archive table; add creation\_date to all tables and automate creation\_date and modified\_date using triggers; add trigger to define geometry\_type; add parameter\_uri and units\_uri to Parameters table; add instrument\_uri to Instrument table; implement honorific\_id in People table to replace title column in People table; change order of columns in Tracking\_status table. For a time, the title column will remain in the People table, but it will be removed once all code is updated.

37. July 26, 2010. Add dataset\_id to the Archive table picture. It was in the table but left out of the picture. Add dataset\_platform\_id to the Archive table to deal with the situation where there is no dataset\_url specified and we have to rely on the dataset\_platform\_url's. This will also allow us to track when new data are added to existing, archived Datasets.

38. August 16, 2010. Add handle to Dataset table. Add comment to People table.

39. August 30, 2010. Fix spelling of graphable in the Parameters table. It had been spelled as graphical. Rename parameter\_uri to parameter\_external\_identifier and units\_uri to units\_external\_identifier in the Parameters table and instrument\_uri to instrument\_external\_identifier in the Instrument table.

40. February 17, 2011. Add archive\_file\_name to Archive table. We are making a major change to the metadata database schema by dividing the Platform table into two, Platforms and Deployments. (Yes, we know, it should have been done this way in the first place.) At the same time, we are separating out synonym into a separate table; geolocation into an intersection table and a Geolocation table; and coordinated\_deployments into a Coordinated\_deployments table. All have ranks associated with the entries to specify ordering of the connections. Also, several other tables are affected since their foreign keys and platform\_id have to be replaced by deployment\_id. This includes the Dataset\_platform table and the Location table. Also, there is a change in name of dataset\_platform\_id to dataset\_deployment\_id in the Tracking table. The Dataset\_platform table is renamed Dataset\_deployment table. The

change to the Archive table will take effect immediately. However, all other changes will take time to test and implement. The schema diagram was modified to reflect the connections between the Tracking-status table and the Dataset\_deployment, Dataset, and People tables.

41. March 7, 2011. After review, we have made the following additional changes to those made in item 40: added in deployment column in the Deployments table (left out in error); renamed Deployment\_geolocation to Deployments\_geolocation; renamed Coordinated\_deployments to Coordinated\_deployment; added affiliated\_id and platform\_url to the Platform table; renamed Synonyms table to Deployments\_synonym table; added Authority table. Note that the platform\_url points to the specific ship if available, or to the operator's site if not available. While it is possible for the platform\_code to be NULL in the Platforms table, we will try to force it to be unique. The deployment column in the Deployments column should be constrained to be unique (as in the past). Change the Deployments table name to Deployment. References to dataset\_platform\_id are changed to dataset\_deployment in tables Dataset\_deployment, Start\_stop, Tracking\_status, and Archive. The entries in the Lookup table for dataset\_platform\_current\_state\_lookup and platform\_activity\_lookup are changed to dataset\_deployment\_current\_state\_lookup and deployment\_activity\_lookup, respectively. In the renamed Dataset\_deployment table, platform\_activity\_lookup is renamed activity\_lookup and dataset\_platform\_url is renamed dataset\_deployment\_url.

42. April 4, 2011. These are changes to the schema picture: The spelling of the primary key in the Coordinated\_deployment table was changed to coordinated\_deployment\_id; removed the dataset\_deployment\_id column in the Archive table; and the authority\_url name needed to be left justified in the Authority table. In the Deployment table, change the deployment column name to deployment\_name.

43. April 26, 2011. Add the missing line between the Dataset and Dataset\_funding tables. The People\_role table will have the data changed in the table\_name column to reflect the table name change from Dataset\_platform to Dataset\_deployment.

44. May 6, 2011. Add the location column back into the Deployment table for free field input of location Information.



45. May 11, 2011. Fix the diagram, changing the Platforms table name to Platform. Fix the diagram primary key in the Deployment\_synonym table to be deployment\_synonym\_id. Fix the database Start\_stop table column name from dataset\_platform\_id to dataset\_deployment\_id. Add in the authority\_id, affiliation\_id and platform\_url columns in the Platform table. They were left out by accident. Add foreign key constraints to the Platform table and the Start\_stop table.

46. May 26, 2011. Add platform\_title\_lookup to the Platform table since we are splitting out the vessel/platform title from the platform\_name, for example, R/V Knorr is stored as Knorr with the R/V title accesses via the Lookup table.

47. December 14, 2011. Reorganize the way we represent and store award information by creating an award table. The three tables which contained award information, Dataset\_funding, Program\_funding and Project\_funding are renamed Dataset\_award, Program\_award and Project\_award and their corresponding keys are renamed Dataset\_award\_id, Program\_award\_id and project\_award\_id respectively. The award information columns, project\_number, award\_number, and award\_number respectively are replaced by a foreign key, award\_id, to the Award table. Also, the award\_url columns are removed and kept in the Award table, along with new information, program\_manager\_id, a foreign key referencing the program managers name in the People table. Other changes made at this time include: removing the title column in the people table; making the username column in the people table unique; changing the zero values of the honorific\_id column to be NULL; add foreign key constraints to the dataset\_deployment\_id and deployment\_id columns in the Start\_stop table.

48. February 16, 2012: Fix the notes from December 14, 2011 correcting the references from Dataset\_award (etc) to Dataset\_funding. Add platform\_external\_identifier to the Platform table. Also, remove the modification\_history table as it is no longer used. Add in the triggers for the created\_date and modified\_date in the Award table.

49. June 13, 2012: The award\_url column was not shown in the schema diagram as part of the Award table. This was added. The connection between the Dataset and Dataset\_deployment tables was not shown in the schema diagram. This was added. There were no changes to the schema design, only the diagram was corrected.