



Department of Geosciences

Oregon State University

104 Wilkinson Hall • Corvallis, Oregon 97331-5506

Tel: (541) 737-1201 • Fax: (541) 737-1200 • www.geo.oregonstate.edu

December 17, 2007

To: Yassine Lassoued, CMRC

From: S. Mark Meyers, Department of Geosciences, Oregon State University

Oregon State University Principle Investigator: Dr. Dawn Wright

Re: Biological Data Integration (BIDI) Data Loading

Introduction

The purpose of the Biological Data Integration (BIDI) project was to integrate the Marine Institute data into a data structure that unifies all the Marine Institute's data (Cummins and Lassoued 2007). This was accomplished through the blending of the Arc Marine Data Model with the Marine Institute Data Model (Cummins and Lassoued 2007). The outcome of this effort was the development of two BIDI geodatabase (BGDB) model options (see the link to the document *Initial Comments on Various Options* at <http://workshop1.science.oregonstate.edu/fri07>).

As part of their contribution to BIDI, Oregon State University was tasked with loading Marine Institute (MI) sample data into BGDB, Option 2 and reporting on data inconsistencies, incompatibilities, and data loading procedures. Primary data of concern for this task were the Fisheries Surveys and Harmful Algal Bloom data. Sample data arrived in several formats including Microsoft Access databases, Microsoft Excel spreadsheets, and ArcGIS geodatabases. S. Mark Meyers, Senior Faculty Research Assistant, Department of Geosciences, Oregon State University was assigned this task.

Biological Data Integration (BIDI) Model Data Loading

Marine Institute data were determined to be in need of “conditioning,” that is, redefining field formats, creating identification fields or concatenating fields was required. A seemingly bigger problem, however, was with the Access BIO Database and Excel spreadsheets (Microsoft). Particularly true for the Access (Microsoft) databases, the data had become corrupted creating erroneous field entries and field formats. For example, fields were

populated with possible error values such as {0192A2AE-00FB-41E9-B491-68F6374B06BE}. Even if these values were intentional, the format defaulted to text or strings that were not permissible for identification values (ActivityID, OrganizationID) in the BGDB. In addition, the auto cell formatting function in Access and Excel incorrectly formatted fields. That is, fields were formatted as text when they should be integers or interpreted as “Double” precision numbers when they should have been short or long integers when loaded into the BGDB. For some data, fields that should have had integer values actually had text values (e.g., Larvae Biological Data (StationID) which had mostly numeric values also had alphanumeric values which did not appear to be typographical errors).

In general, because of issues with formatting, all tables were removed (exported) from their databases or spreadsheets and saved as either comma delimited (cvs) or DBaseIV (dbf) formats. This cleared the formatting and allowed ArcCatalog to accept the data and apply appropriate formatting when loading. Converting and amending the data in preparation for loading into the BGDB was a two- or three-step process that usually involved Microsoft Excel. Where possible, ArcMap was used to edited tables (OBJECTID required).

Because of the importance of the object classes to the function and integration of the database, the majority of time was spent constructing these object classes starting with the business tables such as Activity, Contact, Country, and Platform. Much of this information was found in the SurveyDatabaseVersion04 database. In the section below those objects (tables) and features classes for which data alignment was possible, data inconsistencies, incompatibilities, and data loading procedures are discussed.

Data Alignment and Corrections

The BGDB object classes (tables) and feature classes were aligned with the Marine Institute sample data (below) to visualize the field-to-field mapping as interpreted. These tables attempt to show where conflicts or incongruities occurred and what corrections were possible.

Object Classes

Activity	FSS_Survey	Alternate Source
-----------------	-------------------	-------------------------

OBJECTID		
ActivityID		ActivityID from SPS_Published_Surveys, I manually matched based on ActivityName
ActivityName	SurveyName	
ActivityDescription	SurveyDescription	
ActivityObjectives		Activity_Objective from SPS_Published_Surveys
ActivityCode		Activity_Code from SPS_Published_Surveys
SPSCode	SPSSurveyCode	
BeginDate	StartDate	
EndDate	EndDate	
ActivityTypeID	SurveyTypeID	
PlatformID	VesselID	
TrackID	Transect	

The Activity object class was derived from FSS_Survey and SPS_Published_Surveys tables found in SurveyDatabaseVersion04 database. Because the Access BIO Database formatted SurveyID as an objectID and dates as text, FSS_Survey was exported to Excel to reformat. Dates had to be reentered by hand in new fields because of mixed formats. Dates were formatted as dd/mm/yyyy, though dates were found in several different formats throughout the MI data.

Activity_ID from the SPS_Published_Surveys appeared as alphanumeric values like this {9A22A6A1-CB01-41B8-888E-A61D5BED0B84}. These values were unique so a new field, ActivityID, was created. The ActivityID field was populated with sequential integer values, a format acceptable to the BGDB. This was an artificial fix and only provided a placeholder. The question remained how to relate SurveyID to ActivityID in a larger context then one cruise.

VesselID required reformatting “Double” to long integer before loading into the BGDB.

The Activity table was completed by hand using data derived from FSS_Survey and SPS_Published_Surveys. Transect was translated to TrackID. For these data, only some transect numbers were located (14, 15, 16, 17, 30, 46, 49, 61, 63, 65, and 66). It was assumed that transects (tracks) were always the same between cruises (values range between 1 and 66).

ActivityType	FSS_SurveyTypes	Alternate Source
--------------	-----------------	------------------

ActivityTypeID	SurveyTypeID	
ActivityTypeName	SurveyType	
ActivityTypeCode		Unknown source

ActivityContactRole	FSS_Scientist	Alternate Source
OBJECTID		
ActivityContactID		Unknown source
ActivityID		Unknown source
ContactID	Contact_ID	
ContactRoleID	Contact_Role_ID	

Sufficient data were not located to develop ActivityOrganizationRole, ActivityDevices, ActivityGears and ActivityMetaData

Contact	FSS_Scientist	Alternate Source
OBJECTID		
OrganizationID	Organization_ID	Value I created in SPS_Published_Surveys
ContactID	Contact_ID	
Surname	Surname	
FirstName	FirstName	
Title	Title	
Address1	Address1	
Address2	Address2	
Address3	Address3	
Address4	Address4	
City	City	
TelephoneWork	Telephone_Work	
FAX	Failed because of mixed formats	
EMAIL	EMAIL	
JobTitle	Job_ID	
CountryID	Country_ID	

FAX failed to load because entries were of different formats (00441502513865, 01-6604462). The BGDB recognized this field as “Double” precision numbers. FAX was not repaired. Creating a new field of sequential numbers generated OrganizationID. This was the substitution for the alphanumeric values ({0192A2AE-00FB-41E9-B491-68F6374B06BE}) found in the SPS_Published_Surveys table. Telephone_MOD and Telephone_Home could be added to the Contact table if required.

Platform	FSS_Vessel	Alternate Source
OBJECTID		
PlatformID	Platform_ID	
CountryID	Country_ID	
OrganizationID	Organization_ID	Value I created in SPS_Published_Surveys
PlatformName	PlatformName	
PlatformCode	Platform_Code	
PlatformCallsign	Platform_Call_Sign	
PlatformTypeID	Platform_Type_ID	

As with most tables the FSS_Vessel table needed to be exported to a comma delimited (.csv) file before loading. Creating a new field of sequential numbers generated OrganizationID. This was the substitution for the alphanumeric values ({0192A2AE-00FB-41E9-B491-68F6374B06BE}) found in the SPS_Published_Surveys table. This was an artificial solution.

Country	Feature Class Europe	Alternate Source
OBJECTID		
CountryID	NATION	
CountryName	CNTRYNAME	
CountryCode	CNTRYABB	

Country was easily extracted for the Feature Class Europe.

Cruise	FSS_CruiseCodeTable	tblCruiseInfo	Alternate Source
OBJECTID			
CruiseID		Cruise_Number	
Code	New_Code		
Name			Undetermined
Purpose			Undetermined
Status			Undetermined
Description			Undetermined
StartDate			Undetermined
EndDate			Undetermined

ShipName		Ship	
----------	--	------	--

The Cruise table was partially constructed from tblCruiseInfo and FSS_CruiseCodeTable by hand. Only matches for the CruiseID, Code, and ShipName fields were found. Dependent on the assumption that cruise_code = SPSCode.

Species

The Species table was derived from the DeepWater and Groundfish surveys. Species from the HAB were not added to the table. Species names (scientific names) were variable in format. Sometimes subspecies were written with first letter abbreviations (H-D-DENDROCHOROTA). This was acceptable if H and D were previously known. In some cases a family name (e.g., HIATELLIDAE, JANIRDAE) or order name (e.g., HYDRODIA) was present in the *species* name field. It is understandable that individuals cannot be identified to genus or species; however, this type of entry should be treated differently. The species information was loaded into the Species object class in the BGDB. MAFF and NODC fields were added to the table. SpeciesID was created in ArcMap. However, this was an arbitrary value. Species ID probably already existed in either an MI taxonomic scheme or an international scheme and should be provided by the MI. The genus, family, order, and kingdom tables were not practical to develop here.

Marine Features

Track

In general tracks can be derived from FSS_Transects. Because data for transects (tracks) were not collected continuously for duration of the track but collected in short time spans, only segments were developed directly from the FSS_Transect table. Creating the complete paths for tracks was not pursued.

Track	FSS_Transects	Alternate Source
OBJECTID		
FeatureID		
FeatureCode		
StartDate	missing	
EndDate	Date_E	
VehicleID		
CruiseID (integer)	Cruise_Code (text)	Depends on development of Cruise table

TrackID	Transect	
Name		?
Method		?
Description		?
LocalDescription		?
Depth	Depth	added
PRC_NASC	PRC_NASC	added
Region_cla	Region_cla	added
Strata	Strata	added
Layer	Layer	added

Dates in FSS_Transects were entered without separators (/ or -, yyyyymmdd) and appeared as double precision numbers to the geodatabase. This was corrected by hand in Excel. The Track feature class was developed but few of the attributes were compatible with the geodatabase. Date_S (StartDate) was missing from the FSS_Transects data. Extending the Track feature class by adding attributes (Depth, PRC_NASC, CruiseID, Strata and Layer) required a decision on which attributes to include and which to exclude. This decision should not be made by a data manager or data entry technician only by the data owner(s).

ICEDivisions

The ICEDivisions feature class was derived from the ICE_Divisions feature class found in NDP Project Egg Larval database. The ICE_Divisions properties in the BGDB showed ICECODE as a long integer. In the MI sample data, ICES_Divisions, ICECODE is a text field. The field was actually populated with letters, Roman numerals and Arabic numerals (e.g., IVa, Vb1, X). The only field match possible is the ICESNAME. ICESREC_ID may be a reasonable substitute for ICECODE but in other places in the data sets where ICESREC_ID was used the field was populated with Roman and Arabic numeral (e.g., IVa, Vb1, X) values (see FSS_HAULS). In the Stations data ICESCode values appeared completely different (27E3E, 26E2E). Definitions and translations are needed.

ICESDIVISIONS	ICES_DIVISIONS	Alternative Source
Shape	Shape	
OBJECTID	OBJECTID	
FeatureID		
FeatureCode		
ICESID	ICESREC_	
ICESCode	ICESREC_ID	
Name	ICESNAM	

Shape_length	Shape_length	
Shape_Area	Shape_Area	
ICES_Code	ICESCODE	added
ICESREC_ID	ICESREC_ID	added

Fields Q1, Q2, Q3, Q4, Total, Log1, Log2, Log3, and Log4 were dropped.

Egg and Larval Data

The InstantaneousPoint feature class seemed appropriate for LarvalBiologicalData. Generating the points from comma-delimited or dbase files avoided Excel formatting errors. Larval biological data dates were in three separate fields: day, month, and year. These three fields were concatenated in Excel to one field (dd/mm/yyyy) before being exported. It was uncertain what to do with StationData (StationsLayer). By definition these data belong to the InstantaneousPoint feature class.

SeaZoneOffshoreDepthContours

This data set matched the definition of FeatureLine as it was derived from depth measurements. It was loaded without change.

HAB/BioAssyChemistry

The text found in Sample_Code (BTX05022002) looked like a date. Should this be an integer field? Text values found in the result field were of the form: values = negative, <LOD, <LOQ. The BGDB was looking for an integer format.

Europe

This feature class was imported into the geodatabase as marine feature class Europe for visual reference.

Comments

It was difficult to compile all the data necessary for the tables without better knowledge of the sources and without an instructive guide. Therefore, field-to-field mapping guides for each object class and feature class should be created in the future. Such guides would include data relationships and definitions. This would help avoid making unsubstantiated assumptions about the data or fundamentally altering the meaning of the data.

The data owners (MI) should be involved in all data manipulation decisions to avoid fundamentally altering the meaning of their data. Though it was easy to create various identification fields such as SpeciesID it may not have been appropriate.

It would be ideal to work with the raw data and not data filtered through Microsoft Access or Excel. Data manipulation should probably be done via scripting using an appropriate language such as Python (<http://www.python.org>) or Perl (www.perl.org) For their Marine Data Repository, the MI uses Microsoft SQL Server, ArcSDE and Python scripting for extracting, transforming and loading (Wright et al. 2007, p. 83). In addition, these scripts keep track of FeatureIDs and MeasurementIDs which were missing from the MI sample data. Perhaps similar methods could be applied for the BGDB.

References

Cummins, V and Lassoued, Y. 2007. *Biological Data Integration: Data Design Options*. NDP Marine RTDI Strategic Programme.

Wright, D. J., Blongewicz, M. J., Halpin, P. N. and Breman, J. 2007. *Arc Marine: GIS for a Blue Planet*, ESRI Press, Redlands, CA.

BIDI Geodatabase Model Option 2 (BGDB)

Data loaded into the BGDB may be found at this FTP site:

<http://my.science.oregonstate.edu/~meyerss/outgoing/>. Just click on

AcousticSurvey2.2a.zip and save to disk when pop-up appears (MS Windows XP). The data are also mirrored at <http://workshop1.science.oregonstate.edu/fri07>.