Exercise 7 – Creating a Relational Database from a Messy Dataset

Session Objectives: At the conclusion of this session, you will be able to:

- ❑ Clean-up a long-term monitoring and inventory dataset using ArcGIS and MS Access.
- ❑ Restructure a flat table into a relational database.
- ❑ Use MS Access to view, query and analyze data in related tables.
- ❑ Connect to an MS Access database and import tables into a geodatabase.

Material Created By: Eric Kelchlin (November 2013)

Revision: March 2014

Software: ArcGIS 10, SP5

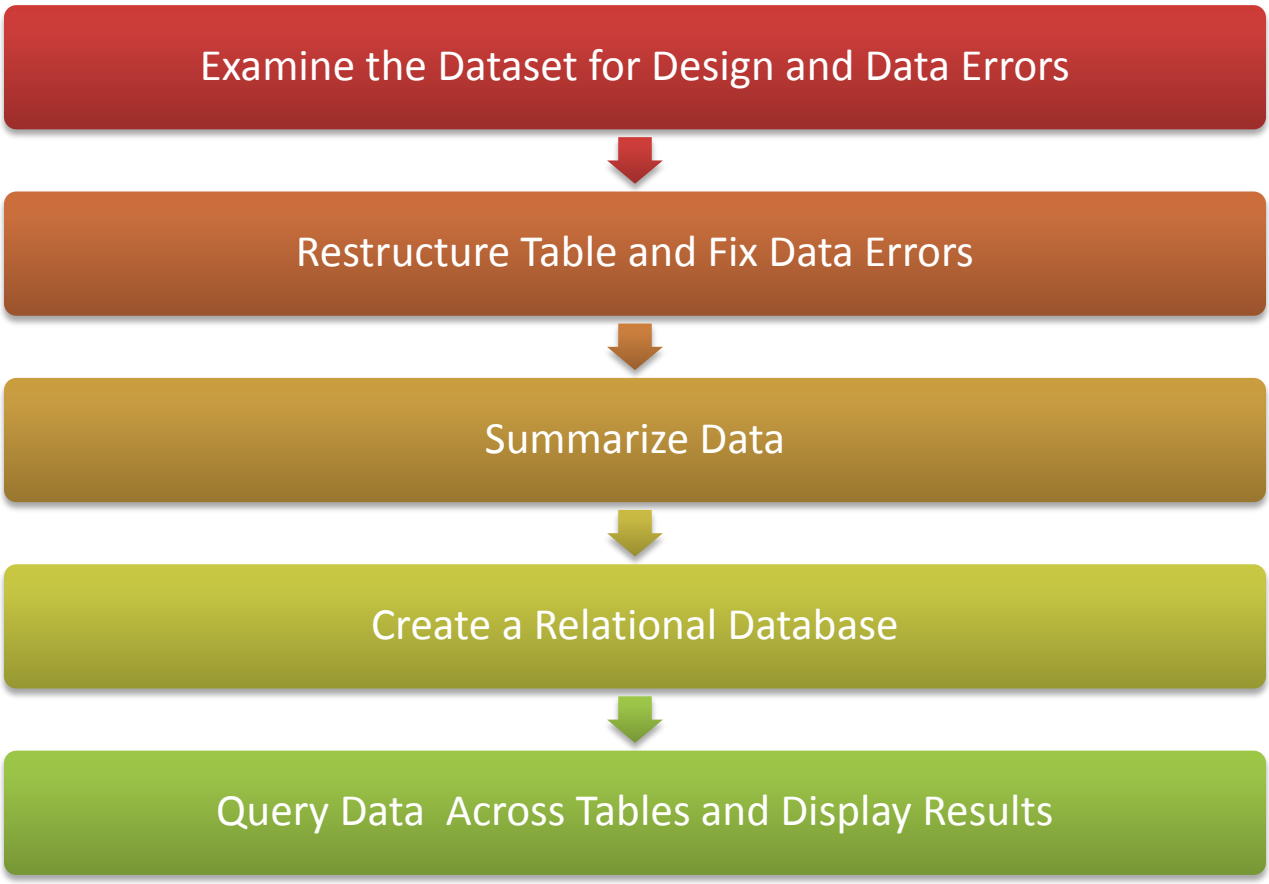
Data Path: C:\ or D:\CSP7200\Exercise 7

Databases: Bombayhook.gdb

Tables: Plants.dbf

MXD: Exercise7.mxd

Workflow



Data Management

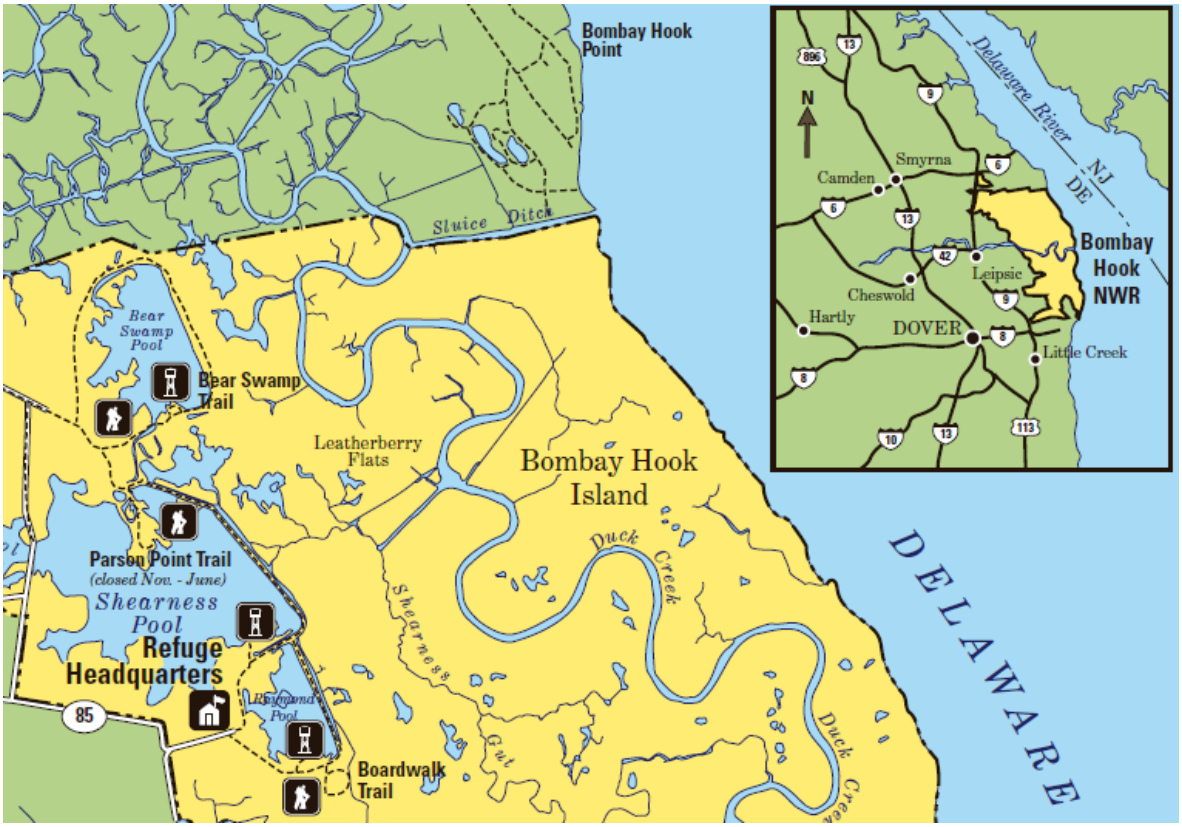
In this Exercise, you will utilize ArcGIS to clean-up a 15 year long wetland monitoring dataset from Bombay Hook National Wildlife Refuge. The dataset contains location information on sample sites (n=220), saltmarsh vegetation species and percent cover values that are expressed as Daubenmire cover class mid-points.

Cleaning-up messy data involves a systematic process of finding errors and fixing them. The objectives are project specific but they typically involve searching for and fixing typing errors, deleting records that are redundant or useless for data analysis, merging similar records (this is no time to be a splitter!), recoding values and restructuring the table for data analysis.

Long-term monitoring and inventory datasets are plagued by errors because the sampling protocol and sampling intensity changes annually due to employee turnover and short-sighted funding. Observer bias can also be a major problem with long-term data sets. Some field technicians are more skilled at plant and animal identification than others, which results in a dataset rich with species diversity in some years and just deplorable in other years.

The goal of this project is to monitor salt marsh plant species composition and percent cover changes though time. **Our goal is to** clean-up the dataset, restructure the flat table into a relational database with multiple related tables and examine plant species trends.

Study Site



TASK 1 – Examine the Dataset for Design and Data Errors

All data are located in the **CSP7200\Exercise7** folder.

The first step is to explore the dataset(s) to find errors, redundancies, missing values and determine if the table structure can be reduced and broken-up into multiple tables.

1. Open the **Exercise7.mxd**.

2. Open the Attribute Table and examine the dataset for errors.

   - Do you see any potential issues with the dataset?
   - Should we simplify the field names?
   - Should we restructure the dataset into multiple related tables?

3. What fields represent the following variables:

   - Sample Sites_____
   - When Sampled_____
   - Plant Names_____
   - Sampling Unit (i.e., percent cover estimate)_____

TASK 2 – Table Restructure

The biologists visited each sample site an annual bases between 1991 and 2007. A GPS position was retaken each time to create a record for data entry. The result is a stack of points on top of each other. I would like to reduce this stack down to the total number of unique sample sites. This dataset will become the primary table that all the other secondary tables are linked to via a series of relationship classes. This table will also be the only one that contains the geometry of the point feature; all other tables will contain attributes that are linked to the primary table using the unique sample site name.

A first step would be to restructure the dataset into multiple tables starting with the sample sites field (**Im_Biol_Monit_Sites_Pt_Pt_Name**). To do this, let's first dissolve

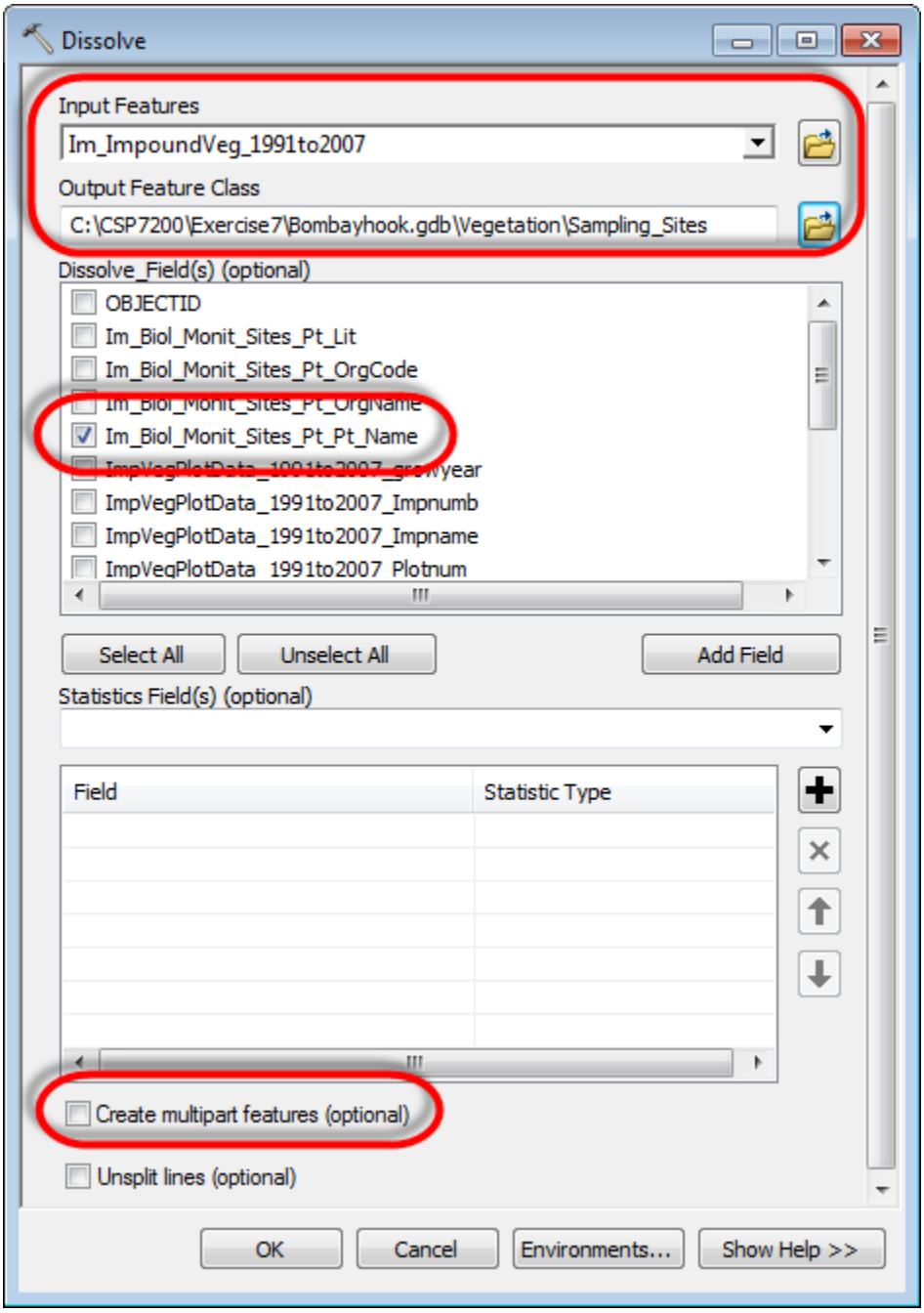
1. Dissolve the table by sample sites field.

   - Click on the **Geoprocessing** toolset from the main toolbar and select **Dissolve**.

- Name the new feature class **Sampling_Sites** and save it in the **Vegetation** feature dataset within the Bombayhood.gdb.

- Use the **Im_Biol_Monit_Sites_Pt_Pt_Name** at the dissolve field as shown below.

- Make sure the Create Multipart Features box in the bottom of the dialog is unchecked. Click **OK**. The new feature class will be added to the Table of Contents.



2. Open the **Im_ImpoundVeg_1991to2007** Attribute Table. How many records are in this table _____?

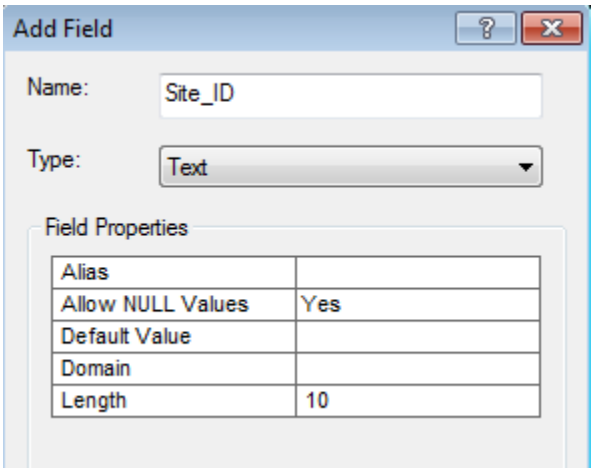
3. How many records are in **Sampling_Sites** _____?

   So, now you can see that for every plant recorded in the field, a separate point was created to add the plant data. This is a very poor database design. The annual monitoring database should be designed as a relational database to avoid taking GPS points each year. You only need to fill-out related tables that are linked to the GPS points.

4. Open the **Sampling_Sites** attribute table to simplify the field names by creating a new field and copying over the data into the new field. You could rename the field directly, but it will only be an alias.

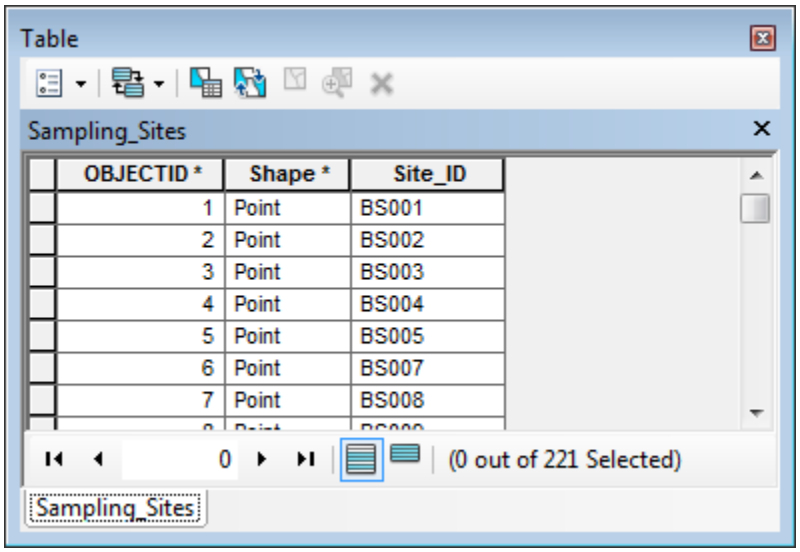
   a. Add a new field from the **Table Options** button .

   - Name this new text field **Site_ID**. Use a length of **10**.



   - Next, right-click on the **Site_ID** field and select **Field Calculator**. Copy the values from the neighboring field by simply double-clicking on the **Im_Biol_Monit_Sites_Pt_Pt_Name** field. Click **OK**.

   - Right-click and delete the old field. Close the table.



TASK 3 – Load Data into a New Table and Export to MS Access

Many of the fields that are not necessary for our purposes and the field names are cumbersome. This section shows you how to load (i.e., extract) only the data we need and place it into a new table. This method does not affect the original dataset and is therefore much safer to making erroneous errors than can't be undone. Let's load only the data we want into a table and fix many issues in one step.

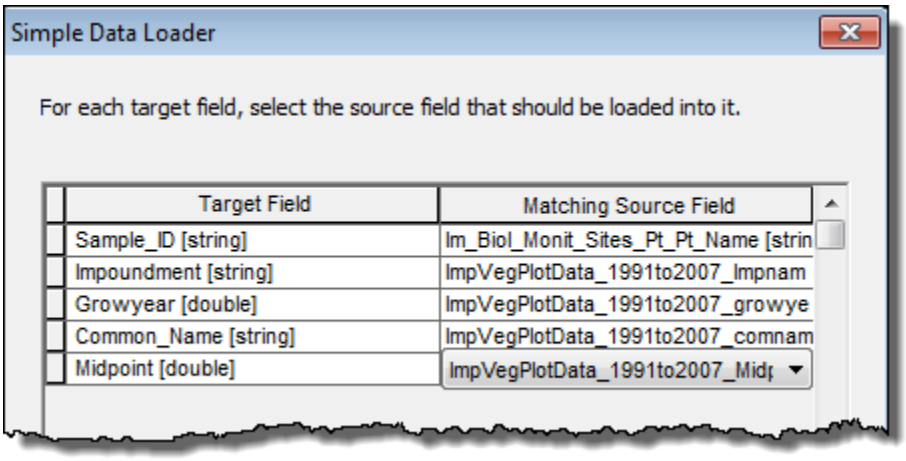
1. Open ArcCatalog from the docked sidebar.

   - Right-click on the **Bombayhook** geodatabase and view the **Data1** table structure. The table structure is shown below as well. We've already created the table for you to save time.

Field Name	Data Type	Text Length
Sample_ID	Text	10
Impoundment	Text	35
Growyear	Double	
Common_Name	Text	35
Midpoint	Double	

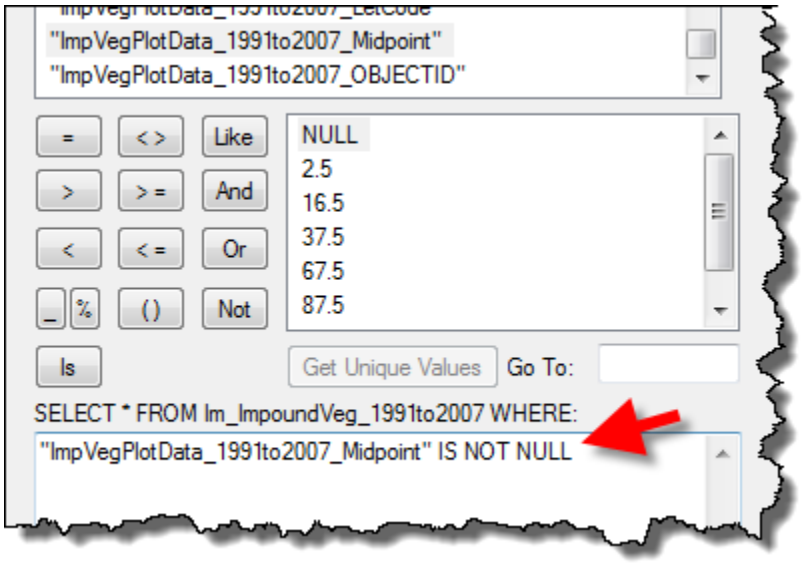
2. In ArcCatalog, right-click on the **Data1** table and select **Load Data**.

   - Use the **Im_ImpoundVeg_1991to2007** feature class as the input dataset.
   - Match the Source fields with the Target fields, click **Next** once.



   - Build a SQL query

to remove the midpoint records with **Null** values. See the examaple below. Verify the query then click **OK** and **Finish**.



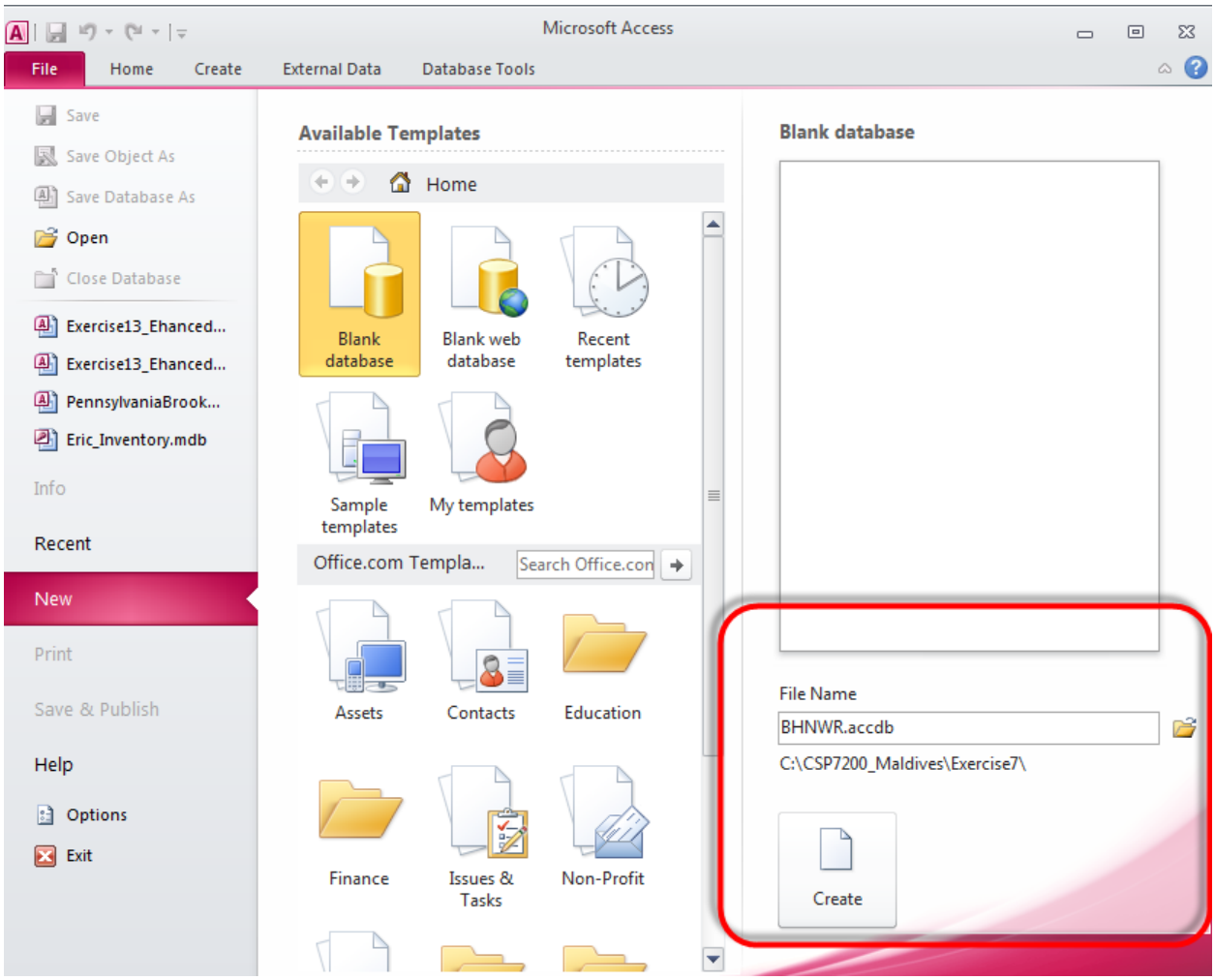
3. Export the table as a .dbf file for import into MS Access. Name the file **Data2.dbf** and save in the Exercise7 folder.

TASK 4 – Import Data into MS Access

1. Launch MS Access and create a new database.

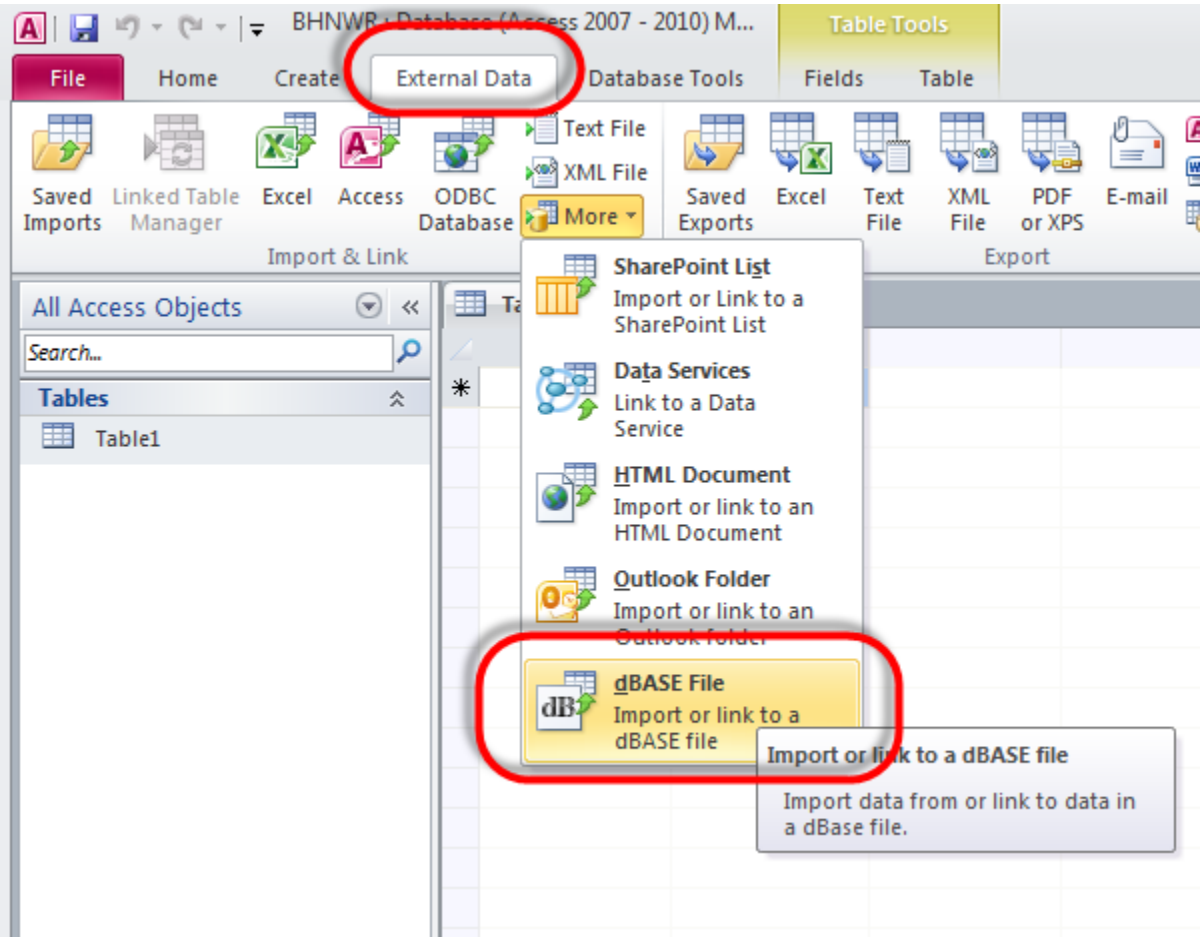
   - In the lower right corner of your screen, browse to our working folder and name the new database **BHNWR.accdb** (short for Bombayhook National Wildlife Refuge).

   - Click the Create button when done.



2. Create a new table from the **Data2.dbf** file you saved from Task 3.

   - Click the External Data tab > More button and dBASE File option.
   - Browse to the location of the **Data2.dbf** file and click **OK** and **Close**.
   - You should now see a table called Data2 in the section of Access called the Objects Navigational Pane.

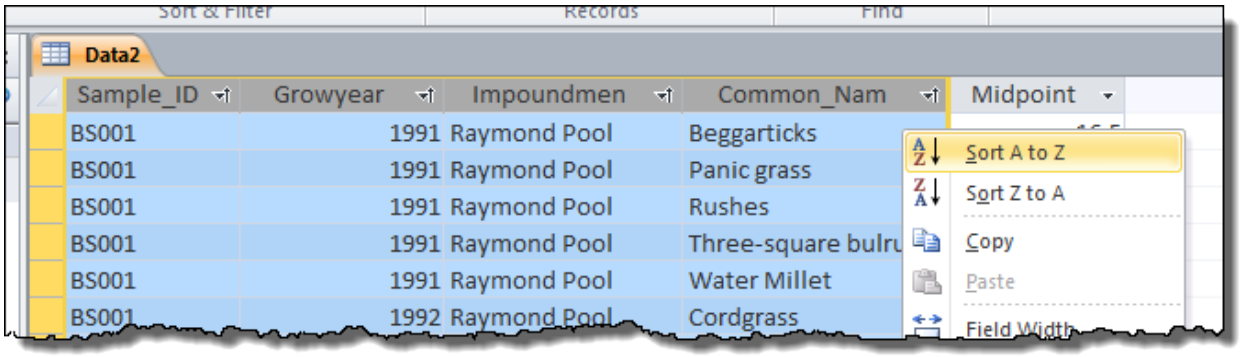


TASK 5 – Summarize the Data

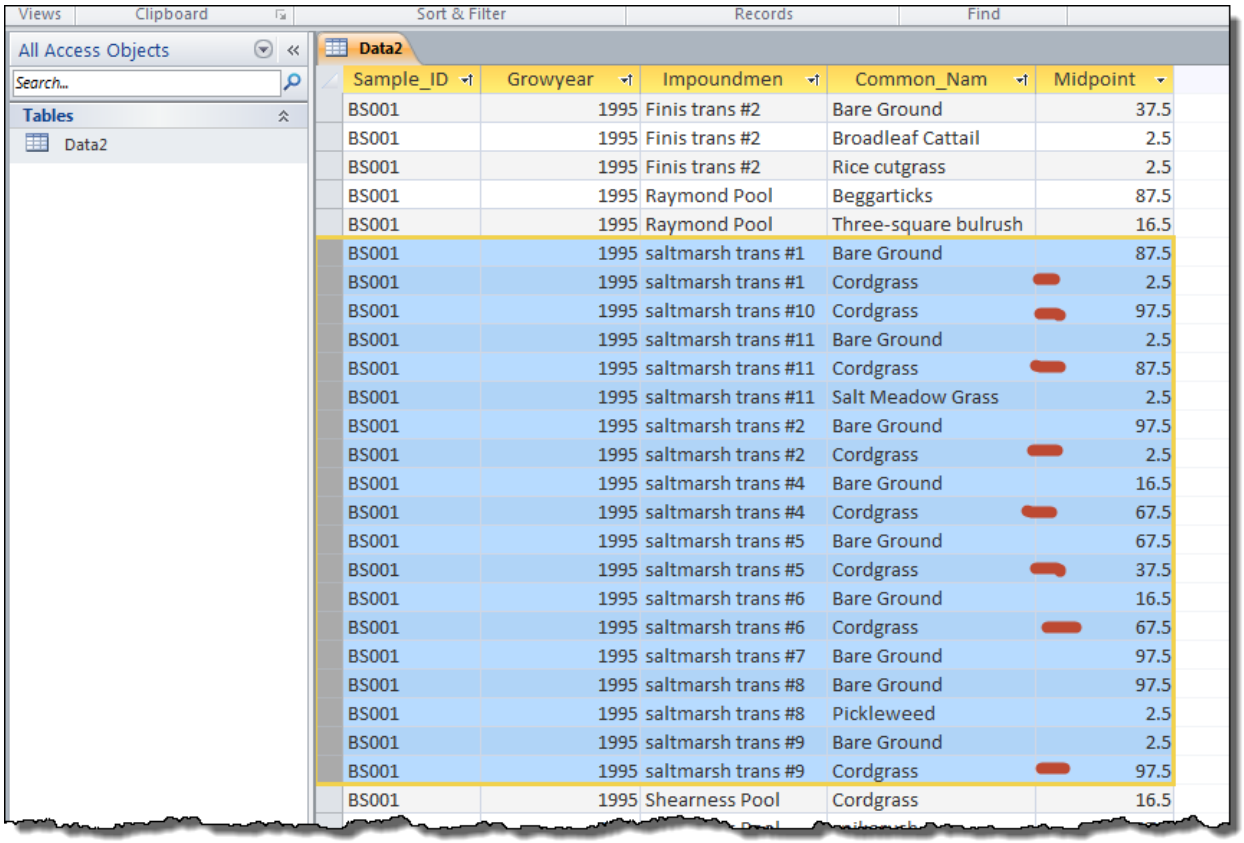
Multiple samples/transects were taken at the same sample sites in some years. This could be an artifact of different sampling protocols, but we really don't know why. The best we can do is calculate a mean midpoint value for each sample site, relative to year and plant species.

1. Sort the records in the Data2 table across multiple fields at once.

   - Double-click on the Data2 table to view the table in the data view.
   - Move the Growyear field before the Impoundmen field by left-clicking on the header, holding down the mouse button and dragging the field over. This does not change the design of the data table; only how we view and sort the data.
   - Next, select all the fields except the midpoint field by selecting the Site_ID header and dragging your mouse over to select all the other fields.
   - Right-click on the highlighted area and select **Sort A to Z**. You can also find the sort on the Home tab section of the main menu.



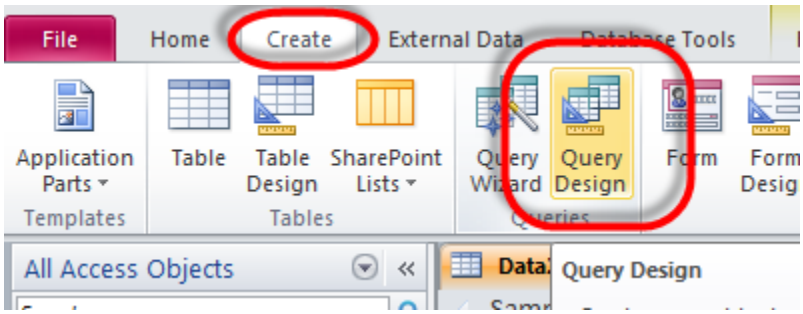
2. Browse down the table and notice how the dataset contains multiple subsamples (i.e. transects) for some of the sample sites within each year. The dataset has should be summarized by sample site, year and plant species before we proceed to any analyses.



Sample_ID	Growyear	Impoundmen	Common_Nam	Midpoint
BS001	1995	Finis trans #2	Bare Ground	37.5
BS001	1995	Finis trans #2	Broadleaf Cattail	2.5
BS001	1995	Finis trans #2	Rice cutgrass	2.5
BS001	1995	Raymond Pool	Beggarticks	87.5
BS001	1995	Raymond Pool	Three-square bulrush	16.5
BS001	1995	saltmarsh trans #1	Bare Ground	87.5
BS001	1995	saltmarsh trans #1	Cordgrass	2.5
BS001	1995	saltmarsh trans #10	Cordgrass	97.5
BS001	1995	saltmarsh trans #11	Bare Ground	2.5
BS001	1995	saltmarsh trans #11	Cordgrass	87.5
BS001	1995	saltmarsh trans #11	Salt Meadow Grass	2.5
BS001	1995	saltmarsh trans #2	Bare Ground	97.5
BS001	1995	saltmarsh trans #2	Cordgrass	2.5
BS001	1995	saltmarsh trans #4	Bare Ground	16.5
BS001	1995	saltmarsh trans #4	Cordgrass	67.5
BS001	1995	saltmarsh trans #5	Bare Ground	67.5
BS001	1995	saltmarsh trans #5	Cordgrass	37.5
BS001	1995	saltmarsh trans #6	Bare Ground	16.5
BS001	1995	saltmarsh trans #6	Cordgrass	67.5
BS001	1995	saltmarsh trans #7	Bare Ground	97.5
BS001	1995	saltmarsh trans #8	Bare Ground	97.5
BS001	1995	saltmarsh trans #8	Pickleweed	2.5
BS001	1995	saltmarsh trans #9	Bare Ground	2.5
BS001	1995	saltmarsh trans #9	Cordgrass	97.5
BS001	1995	Shearness Pool	Cordgrass	16.5

3. To do this, we'll need to create a query.

- Click on the Create tab on the main menu and select Query Design.

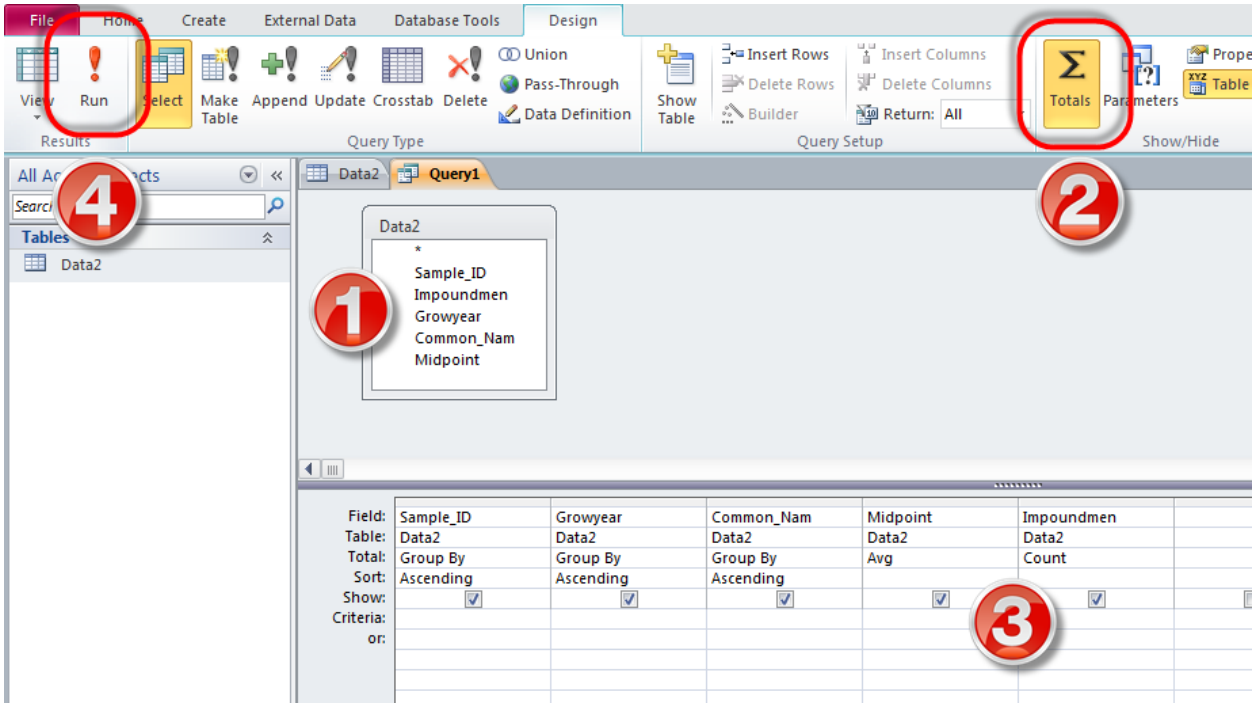


- Click the Add button to bring in the only table we have.

4. To calculate the average midpoint follow these steps:

   - Step 1- Double-click on each field in the Data2 table to add it to the query. Add the Sample_ID, Growyear, Common_Nam, Midpoint and Impoundmen fields.
   - Step 2 - Click on the Totals button

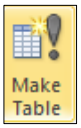
.
   - Step 3 - Change the Group By in the Midpoint Totals row to Avg and to Count for the Impounden field.
   - Step 4 - Click the Run button.



5. How many records do you have now_____?

6. Since this is our clean, summarized dataset that is ready for analysis, let's go ahead and make a new table out of this query.

7. Click the Design View button

and the Make Table query button 

.

8. Name the table **Data_Sum** and click OK.

9. Run

the query and select Yes to accept your changes.

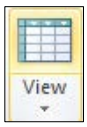
10. Right-click on the Query1 tab and select Close. Click No to save changes. We have our results in the **Data_Sum** table, so there is no need to save the query.

11. Right-click (or double-click) on the **Data_Sum** table to open it. Click on the **Home tab > Design View** button and lets clean the field names up a bit and add some descriptions.

12. Change the field names and add the following descriptions as shown below:

Data2 | Data_Sum

Field Name	Data Type	Description
Sample_ID	Text	Unique ID for each sample site
Growyear	Number	The year the sample was taken
Common_Name	Text	Common name of the plant species
Percent_Cover	Number	Average percent cover
Count	Number	Number of subsamples

13. Click on the Home tab > Datasheet View button 

and save your changes. Note: you can also switch between design and datasheet view from the tiny little icons on the right bottom of the screen.

TASK 6 – Create a Relational Database

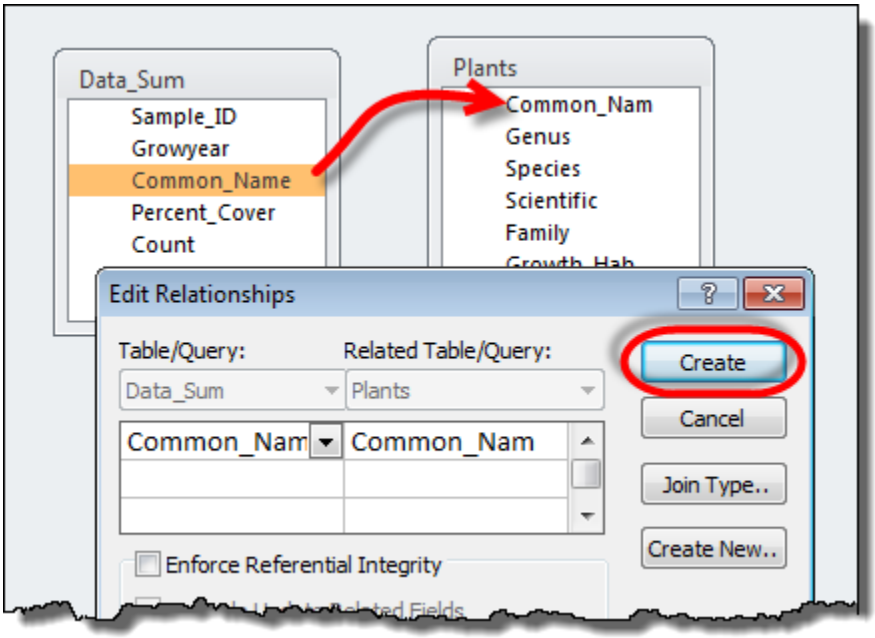
Creating relational databases is all about linking objects (i.e., tables and queries) together using a common field for the purpose of viewing or extracting data across multiple datasets. This design strategy creates a stair-stepped hierarchical data structure, which is in stark contrast to the "flat" mega-tables that scroll-on forever.

1. First, let's import another dbf file.

   - On the main menu select **External Data > More > dBASE File**.
   - Browse to the location of the **Plants.dbf** file and import it into our database.
   - Name the table **Plants**.

2. Next, we're going to create a relationship to link the **Data_Sum** table (the parent/primary table) to the **Plants** table (the child/secondary table).

   - On the main menu select **Database Tools > Relationships** button.
   - Add the **Data_Sum** and **Plants** tables to the screen. Either right-click to add the tables or simply drag them over from the Objects Navigation Pane.
   - To create the relationship, select the **Common_Name** field in the **Data_Sum** table and drag it over to the **Common_Nam** field in the **Plants** table. Click Create.

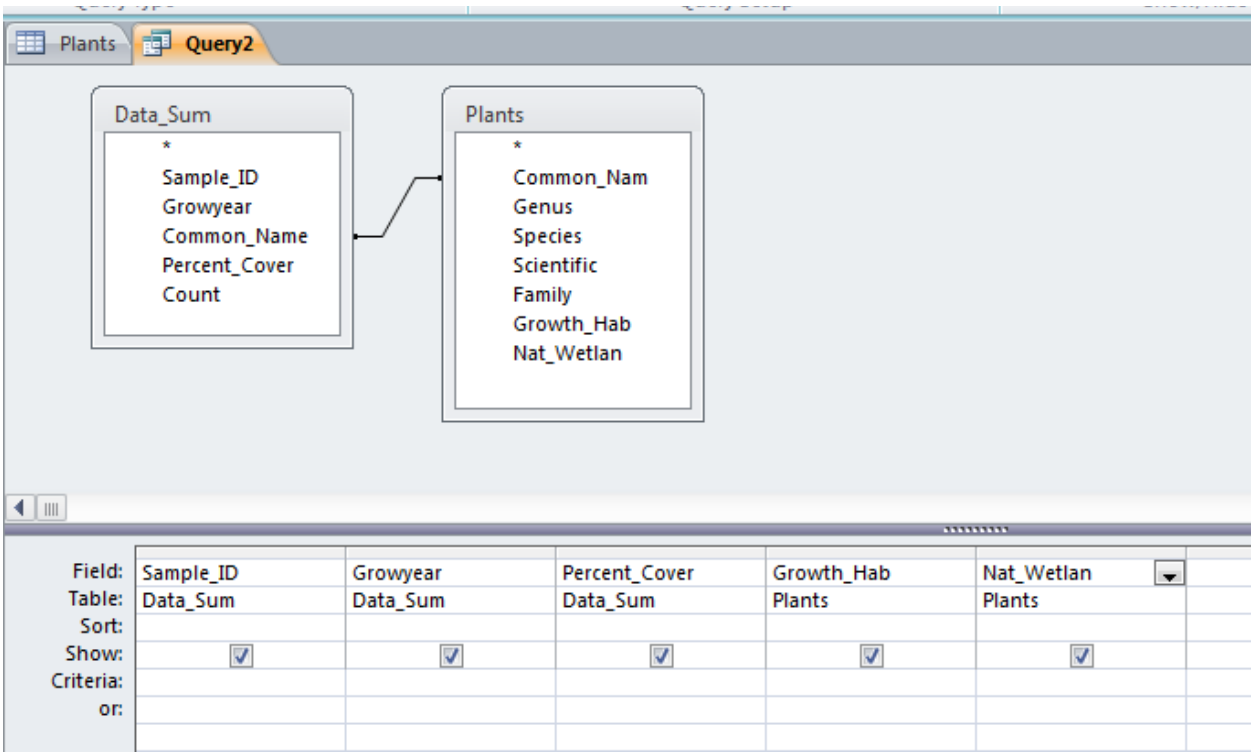


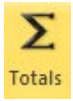
TASK 7 – Query Data from Related Tables and Export Results

Let's say for example that you want to calculate an average percent cover estimate for Graminioids (i.e., grasses) with OBL national wetland status (i.e., obligatory wetland plants) in grow year 2000 for each sample site. Here are the process steps:

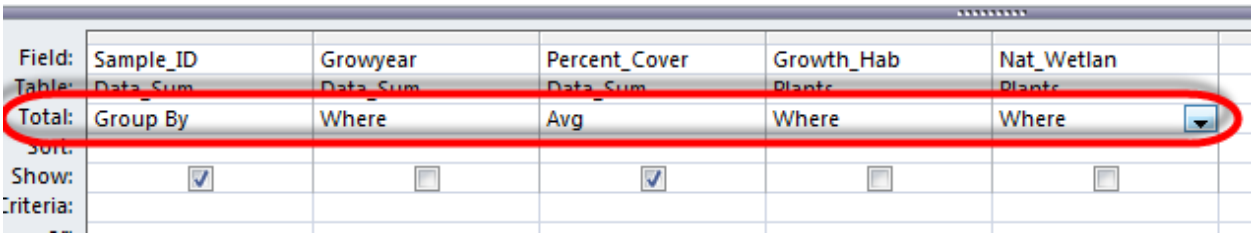
1. Create a new query and add the two tables we just related to each other.

   - Add the fields as shown below.

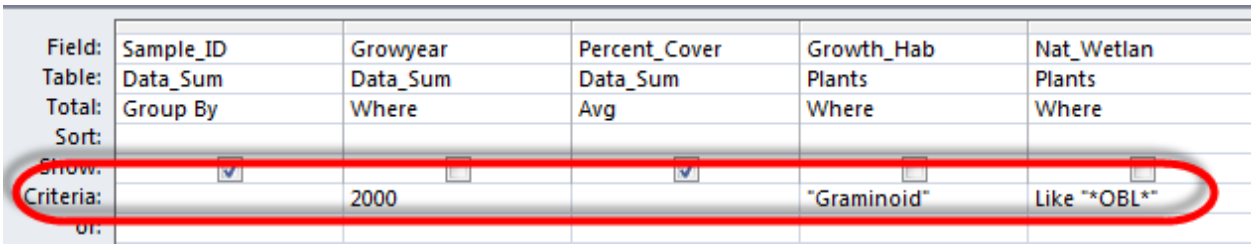


   - Click on the Totals button  to add the Group By totals field.

   - Change the Totals Group By items to the following by clicking the Group By and selecting a different action from the dropdown menu.



- Next, narrow the results by our search criteria. Add the following text in the Criteria field as show below.

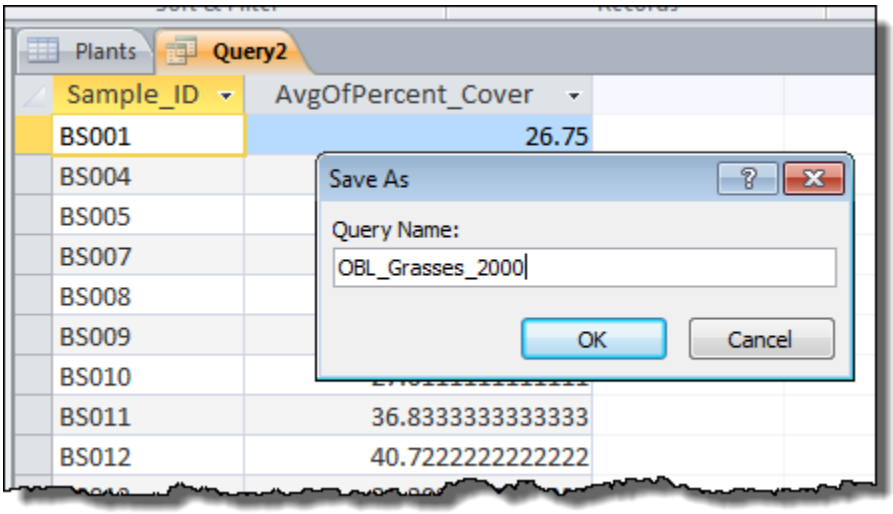


Field:	Sample_ID	Growyear	Percent_Cover	Growth_Hab	Nat_Wetlan
Table:	Data_Sum	Data_Sum	Data_Sum	Plants	Plants
Total:	Group By	Where	Avg	Where	Where
Sort:					
Show:	☑	☐	☑	☐	☐
Criteria:		2000		"Graminoid"	Like "*OBL*"
or:					

- Why do you think we created the **Like "*OBL*"** query SQL language in the Nat_Wetlan field and did not just use **OBL**? Open the Plants table and examine the data to see why.

- Run the query.

- Save the query by right-clicking on the Query2 tab. Select Save and name it **OBL_Grasses_2000**.



2. Let's export this table for use in ArcMap.

- From the main menu, select the **External Data > More > dBASE File**.

- Save the .dbf file in the working folder and exit Access.

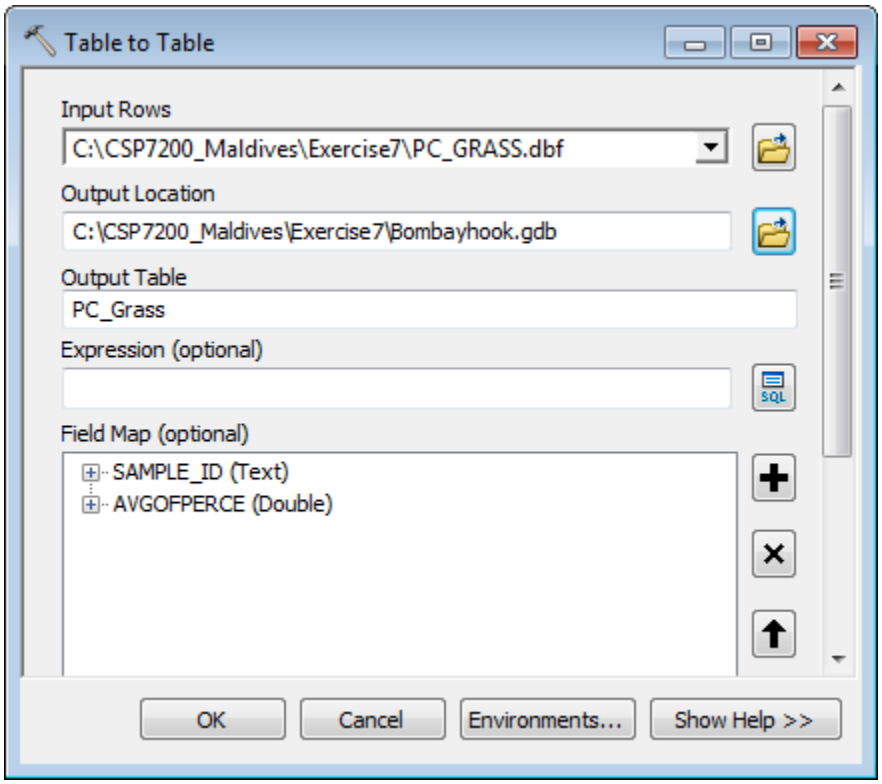
Note: Instead of exporting the results of the query into a .dbf file, you could have linked the query into the Bombayhook.gdb using an OLE DB connection in ArcCatalog. That way any changes you make in Access will show-up in your geodatabase.

TASK 8 – Graphically Display Results in ArcMap

1. Re-open the Exercise7.mxd.

2. Bring in the **OBL_GRASS.dbf** into the **Bombayhook.gdb**. Notice that the file name was shortened from the original OBL_Grasses_2000 file name. Dbf file name can be no longer than 8 characters, so it was automatically shortened from the MS Access Export.

   - From ArcCatalog on the sidebar, right-click on the **OBL_GRASS.dbf** file and select Export. Follow the Table to Table dialog.

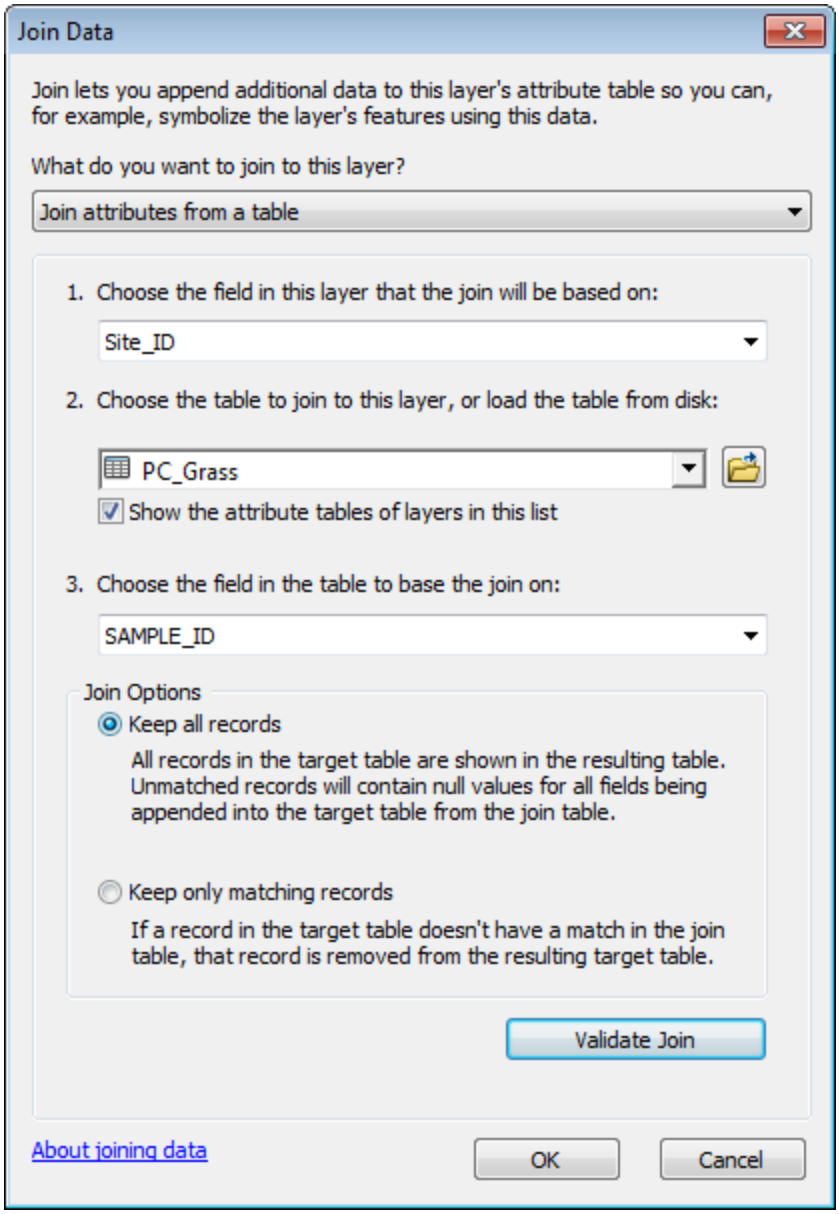


3. Join the data from the **Sampling_Sites** feature class to the **PC_Grass** table.

   - Add the **Sampling_Sites** feature class to the Table of Contents. Remove any other feature classes.

   - Right-click on **Sampling_Sites** and select Join.

   - Fill-in the fields as shown below.

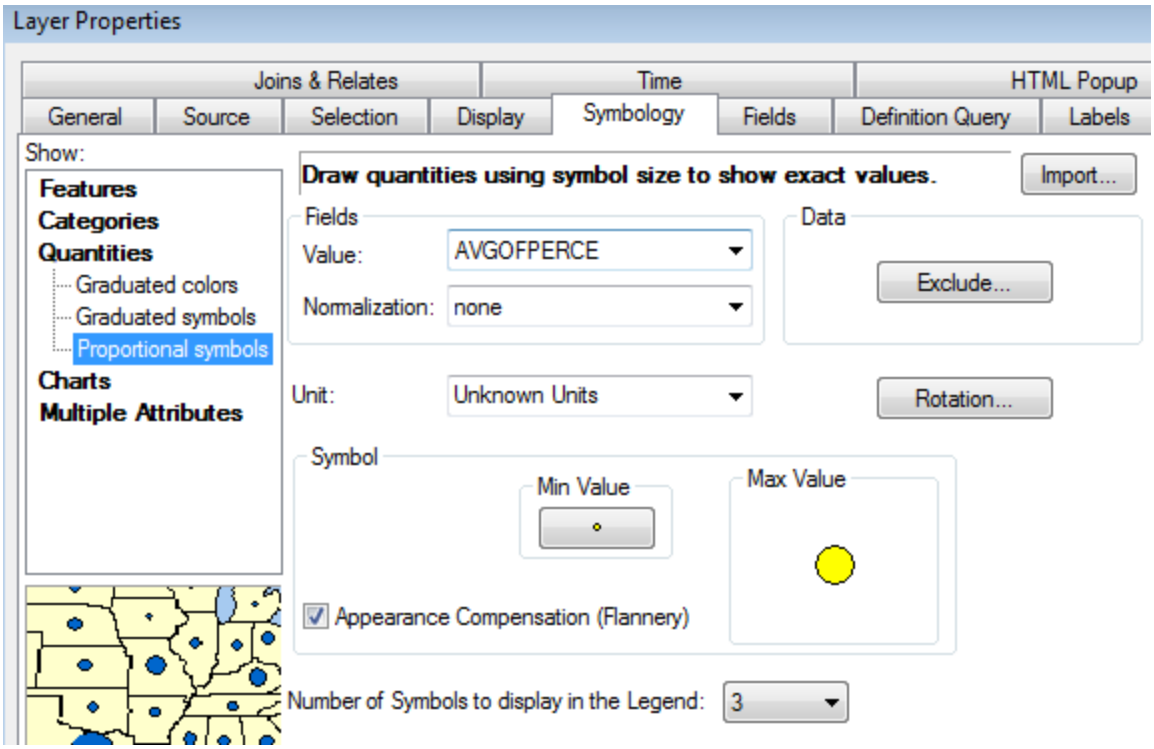


- Click on the Validate Join button to make sure the join will work properly. Often a join will fail because field headers contain invalid characters such as a period "**.**" or a reserved word such as "**Percent**". So it's a good habit always validate you joins.

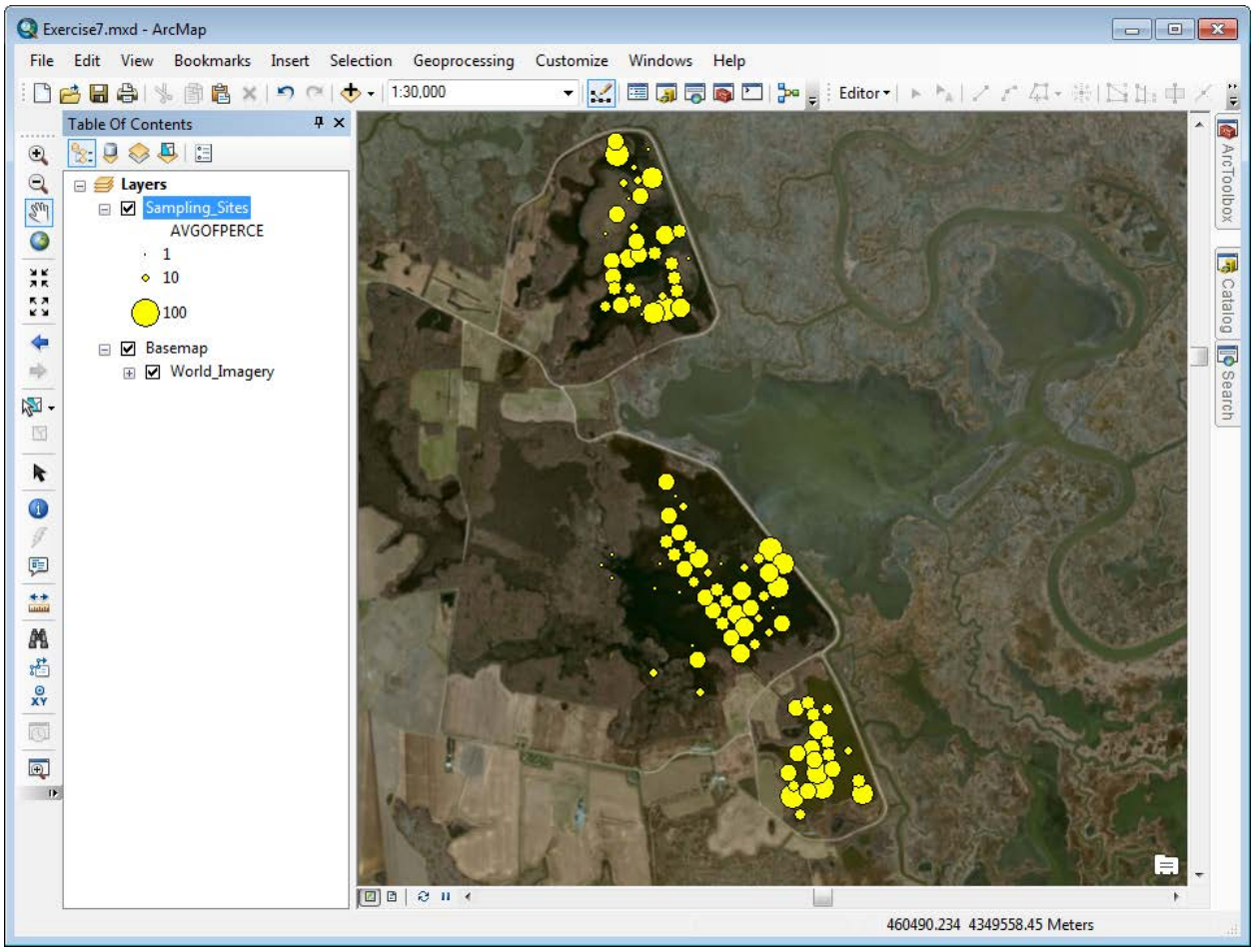
- Click **OK** to complete the join.

4. Finally, let's symbolize each sample point based on the average percent cover value to quickly to visualize trends in wetland obligate grass species from grow year 2000.

   - List your data by drawing order in your Table of Contents.
   - Right-click on the **Sampling_Sites** feature class and select **Properties > Symbology** tab.
   - Change the symbol type to **Proportional Symbols** and use the **AVGOFPERCE** value. Click **OK**.



   - Add a Bing base map to help with your interpretation.


Exercise7.mxd - ArcMap
File Edit View Bookmarks Insert Selection Geoprocessing Customize Windows Help
1:30,000
Editor
Table Of Contents
Layers
Sampling_Sites
AVGOFPERCE
1
10
100
Basemap
World_Imagery
ArcToolbox
Catalog
Search
460490.234 4349558.45 Meters