# X

# *Data management for student projects*

**Exercise 1: Workflow and documentation tools for an undergraduate hydrology project**
Estimated time: 40 minutes

# Exercise 1: Workflow and documentation tools for an undergraduate hydrology project

### Estimated time: 40 minutes

This exercise goes through a short series of data management, geoprocessing, and documentation tasks that typify an undergraduate research project. The first several steps include data preparation and the analysis. The remaining steps focus on exposing and utilizing ArcMap tools to support the documentation process.

The research project (hypothetical) focuses on measuring soil pH in a generally dispersed pattern in a moderately large study area (approximately 63 $km^2$). Soil pH has a direct effect on the availability of macronutrients: nitrogen, potassium, and especially phosphorus. This particular student project will try to show the extent to which pH values are clustered in the study area.

The goal of documentation is the ability to produce identical results from the same source data. There are different forms of documentation: metadata (addressed broadly in a separate EdUC session), data-dictionaries, and workflow documentation.

Metadata is produced to a standard such as Federal Geographic Data Committee (FGDC) or International Standards Organization (ISO). Even though metadata is assigned to a specific dataset at publication, ArcGIS metadata tools can be used at any time to support the documentation of workflow.

Data dictionaries are variously and sometimes loosely defined. At a minimum, a data dictionary describes a database and may be as simple as a table. Data dictionaries typically describe databases as a whole as opposed to metadata which applies to components within the database, such as a feature class. Data dictionaries may also have a specific functional, programmatic role within a relational database management system.

Workflow documentation captures the events that affect (process and step) the end result in an analysis-type project. Workflow documentation is the focus of this part of the exercise.

## Step 1: Familiarize yourself with windows and tools used

☐ Launch ArcMap, navigate to and open the existing map:
c:\student_data_management\soil_analysis.mxd

The map contains a view of a topographic map from western Washington.

☐ On the Standard toolbar, locate the buttons that open the Catalog window, the ArcToolbox window, and the Search window.

☐ From the Geoprocessing menu, locate the option to open the Results window.

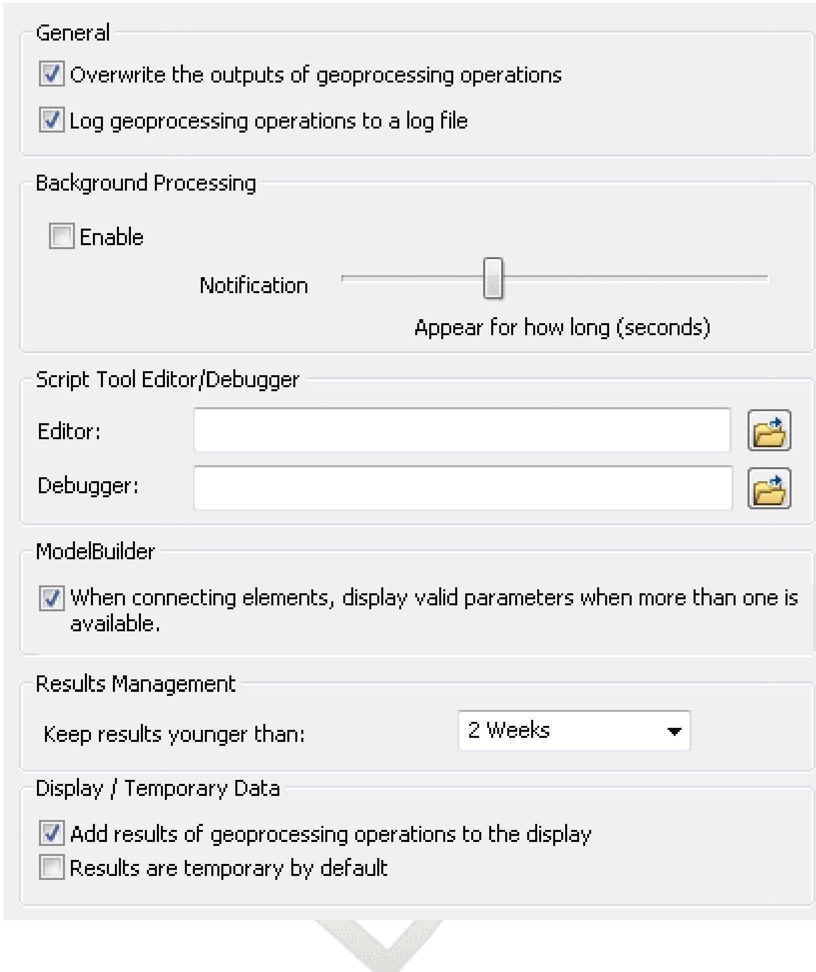
You will use all these windows in this exercise.

☐ Experiment with opening and docking these windows.

As you select and move a window with your mouse, the blue targets in the display indicate docking locations. If the thumbtack is in the vertical position, the window will remain open. If the thumbtack is in the horizontal position, the window will only be open when you mouse over it.

☐ From the Geoprocessing menu, choose Geoprocessing Options.

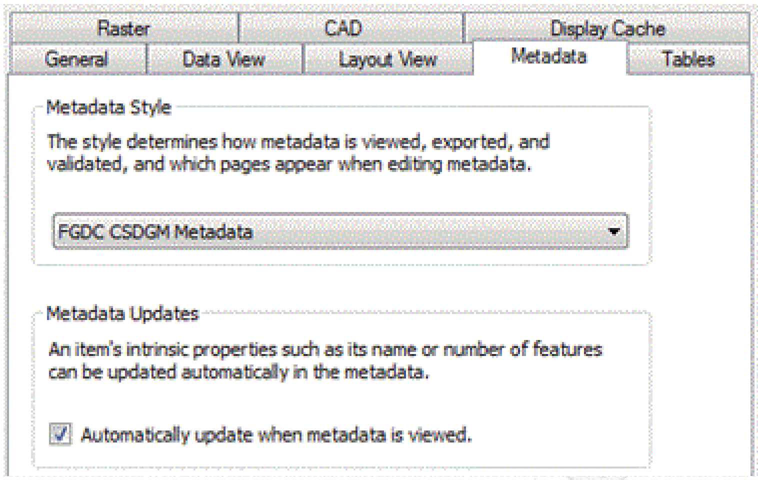
☐ Complete the dialog box to match the following graphic:




☐ From the Customize menu, choose ArcMap Options.

☐ In the ArcMap Options dialog box, click the Metadata tab and set it to match the following graphic:




### *Step 2: Examine field data*

☐ Open and dock the Catalog window.

☐ Expand the Home - Student_data_management directory, if necessary.

The directory contains:

- The file geodatabase, Soil_analysis.gdb
- Soil_analysis.mxd
- Two CSV (comma-separated values) files
- The CSV files contain the output from a GPS data-logger

☐ Right-click the soil_analysis geodatabase and make it the default geodatabase.

It is good practice to associate a default database with any map used for analysis.

☐ Open the tables and review their contents.

Study_area.csv contains longitude and latitude values for points that define the study area. The points were gathered sequentially on the ground and can be used to create a polygon that defines the study area.

Field_obs.csv contains pH values which have been measured at sample points. This is just the first set of multiple samplings that are planned.



## Step 3: Create a boundary polygon that defines the study area from the CSV boundary points

The study area was defined using GPS; however, only a series of points were collected, not a line or polygon feature. Rather than use editing tools, you will use geoprocessing tools to create the study area boundary.

☐ From the File menu, select Add Data and choose Add XY Data.

☐ In the Add XY Data dialog box, browse to and add the study_area.csv file.

The X field and Y field values auto-populate with "longitude" and "latitude," which are the column headings in the CSV file. You need to adjust the Coordinate System of Input Coordinates to reflect the coordinate system of the values in the longitude and latitude columns of the CSV table.

☐ Click Edit.

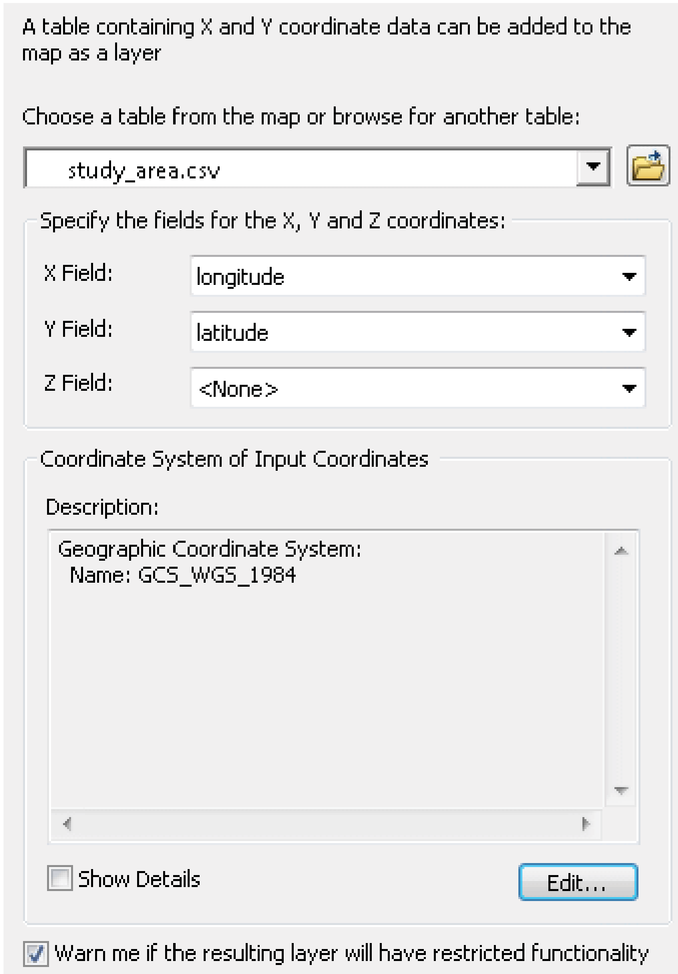
☐ In the Spatial Reference Properties dialog box, click Select.

☐ Navigate to Geographic Coordinate Systems > World.

☐ Select WGS 1984.prj and click Add.

☐ Click OK.

☐ Verify that your Add XY Data dialog box matches the following graphic:




☐ Click OK.

☐ If the Object Id warning displays, click OK.

You can convert the point events—first to a line, then to a polygon.

☐ Open the Search window.

☐ Search for **Points to Line**.

The search returns a number of results.

☐ Choose the script tool, Points to Line (Data Management) and open it by double-clicking the tool name.

☐ Specify the following parameters:

- Input Features: study_area.csv Events
- Output Feature Class: accept the default
- Close Line: check the box
- You can leave the other optional fields blank.

☐ Click OK to run the tool.

☐ Close the Points to Line progress window.

> ***Note:*** In this dataset, the points were collected in the correct order to create a boundary line. If this were not the case, a Sort Field could be created to identify the order in which the points should be used.

The study_area_PointsToLine layer can now be used to create a polygon.

☐ Using the Search window again, find the Features to Polygon (Data Management) tool.

☐ Run the tool with the following parameters:

- Input Features: use the layer that you just created
- Output Feature Class: **StudyArea**
- You can leave the optional fields blank.

☐ Close the progress window when the process completes.

The study area polygon vertices are still stored as geographic coordinates. You will project the study area to the local UTM coordinate system. This accomplishes two things. A standardized coordinate system ensures repeatable results, and spatial statistics tools require projected data.

☐ Search for and open the Project (Data Management) tool.

☐ Populate the parameters with these values.

- Input Dataset or Feature Class: StudyArea
- Output Dataset or Feature Class: StudyArea_UTM

☐ For the Output Coordinate System, click the browse button.

☐ In the Spatial Reference Properties dialog box, click Select.

☐ Navigate to Projected Coordinate Systems > UTM > WGS 1984 > Northern Hemisphere.

☐ Select WGS 1984 UTM Zone 10N.prj and click Add.

☐ Click OK to close all open windows.

☐ Close the progress window.

☐ Symbolize StudyArea_UTM with a hollow fill.

☐ Remove study_area.csv Events, study_area_PointsToLine, and StudyArea from the display.

☐ Save your map.

### Step 4: Create points for all sample sites

Because this process will be repeated every time that sample pH values are measured, it will be more efficient to capture the workflow in a ModelBuilder model.

☐ Open the Catalog window, and expand the Home directory.

☐ Right-click the soil_analysis geodatabase, select New, and choose Toolbox.

☐ Rename the Toolbox **SoilTools**.

☐ Right-click the new SoilTools toolbox, select New, and choose Model.

The ModelBuilder window opens.

☐ From the Model menu, choose Diagram Properties.

☐ In the Diagram Properties dialog box:

- Click the Layout tab.
- In the Orientation area, select Top to Bottom.
- Click OK.

☐ From the Model menu, choose Model Properties.

☐ In the Model Properties dialog box:

- Set the Name value to **SoilpH**.
- Check the box next to *Store relative path names*.
- Verify that the box next to *Always run in foreground* is checked.
- Click OK.

☐ From the Model menu, choose Save.

You will now create the model.

☐ Use the Search window to find the Make XY Event Layer tool, then drag it to the model to add it.

☐ In the model, double-click the tool to open it.

☐ Populate the parameters with these values:

- XY Table: fields_obs.csv (use browse icon)
- X Field: longitude (the tool should recognize the appropriate field names)
- Y Field: latitude
- Layer Name or Table View: accept the default, field_obs_layer
- Spatial Reference: GCS_WGS_1984 (you can navigate to this as you did in Step 3)

☐ Click OK to close the tool.

☐ Use the Search window to find the Feature to Point (Data Management) tool and add it to the model.

☐ Open the tool and specify these parameters:

- Input features: field_obs_layer
- Output Feature Class: pH_measured

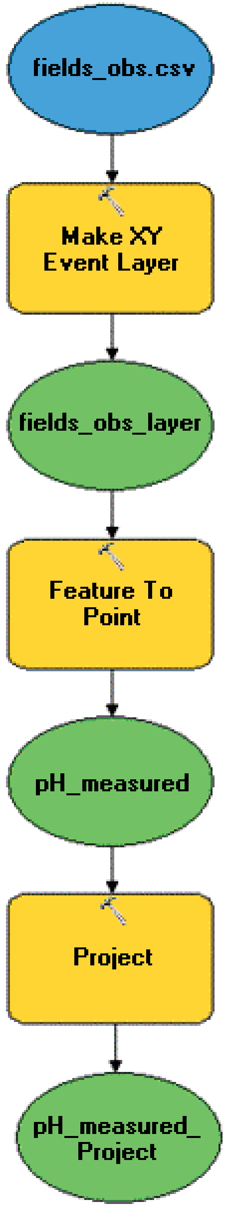
The parameter will recognize the default geodatabase associated with the map.

☐ Use the Search window to find the Project tool and add it to the model.

☐ Open the tool and specify these parameters:

- Input Dataset or Feature Class: pH_measured
- Output Dataset or Feature Class: accept the default
- Output Coordinate System: WGS _1984_UTM_Zone_10

☐ Verify that your model matches the following graphic:




### Step 5: Perform a hot spot analysis to determine if soil pH values are clustered

One of the objectives of this project is to determine the extent to which soil pH values are clustered. There are two components. First you will look at the distribution of the sample sites to establish if they are spatially clustered or dispersed, and then you will perform a hot-spot analysis on the pH values to understand what clustering of pH values is present.

☐ Open the ArcToolbox window.

☐ Expand the Spatial Statistics toolbox and the Analyzing Patterns toolset.

☐ Double-click the Average Nearest Neighbor script tool to open it.

☐ Specify the following parameters:

- Input Feature Class: pH_measured_project
- Distance Method: Euclidean_Distance
- Generate Report: check the box
- Area: **63247028** (the area from StudyArea_UTM)

☐ Run the tool and close the progress window. (You will review the results momentarily.)

☐ In the Spatial Statistics toolbox, expand the Mapping Clusters toolset.

☐ Open the Hot Spot Analysis (Getis-Ord Gi*) script tool.

☐ Run the tool with the following parameters:

- Input Feature Class: pH _measured_Project
- Input Field: pH
- Output Feature Class: accept default
- Conceptualization of Spatial Relationships: Fixed Distance Band
- Distance Method: Euclidean Distance
- Distance Band or Threshold Distance: leave blank
- Self Potential Field: leave blank

To more clearly visualize the results of the Hot Spot analysis, interpolate a raster from the hot spots.

☐ In ArcToolbox, expand the Spatial Analyst toolbox > Interpolation toolset.

☐ Open the Natural Neighbor tool.

☐ Run the tool with the following parameters:

- Input point features: pH_measured_Project_HotSpots
- Z value field: GIZscore
- Output cell size: accept default

☐ Save your map.

A discussion of the Average Nearest Neighbor tool and of Hot Spot analysis is beyond the scope of this short exercise. However, the software documentation on this functionality is comprehensive. Search the help files for **"An overview of the Spatial Statistics toolbox"**.

### *Step 6: Documentation: Results window*

The purpose of this part of the exercise is to expose the functions in ArcGIS that allow the user to capture this workflow for documentation purposes. There are three main ways to do this: the Results window, ModelBuilder Reports, and the ArcGIS component of the metadata document.

The first, and probably the most accessible option is the Results window.

☐ From the Geoprocessing menu, choose Results.

☐ Dock the Results window and pin it in the open position.

☐ Expand Current Session, then expand the Average Nearest Neighbor event.

The results include all the reported values from the progress window, an HTML report, all the parameter values that were used when the tool was run, environment settings, and the full message that displayed in the progress window.

The results of the Average Nearest Neighbor tool can be interpreted from the first set of values. In particular, the NNZScore (Nearest Neighbor Z-score) which is of sufficiently high magnitude to tell us that the distribution of distance values is not normal.

☐ Double-click the HTML Report.

A graphic interpretation of the test opens in a web browser. The results of the test (which compares the distances between points in your data with a set of normally distributed distances for the same number of points and same study area) clearly show that your locations are highly dispersed. This will help you interpret your hot-spot analysis and allow you to infer that clustering of pH values is independent of the distribution of sample sites.

When running the hot spot analysis, you left the distance band parameter blank. For documentation purposes, you need to find the value that the software used.

☐ Find the distance value (*hint*: look in messages) and record the value here: _________________________.

☐ Using the Results window, find and record the output location for the raster that you created with the Natural Neighbor tool. (*Hint*: right-click the output raster.) ________________________________________________________

## Step 7: Documentation: ModelBuilder report

In this step, you will look at the ModelBuilder report tool.

☐ In the Catalog window, expand Home - Student_data_management > soil_analysis geodatabase.

☐ Expand the SoilTools toolbox, right-click your model and choose Edit.

☐ From the Model menu, choose Report.

☐ Choose to view the report in a window, then click OK.

This report, at the highest level, shows Variables and Processes.

☐ Click the link to Expand.

The Model Report shows the model in its current condition. It is an effective way to capture information needed for documentation—to compare the input and output spatial references (perhaps to verify whether or not a geographic transformation was needed).

☐ Expand Processes, Make XY Event Layer, and Parameters to see the input spatial reference.

☐ Expand Project, Parameters, and Output Coordinate System to see the output coordinate system.

☐ Was a geographic transformation required? _________________

No, both use WGS 1984.

## Step 8: Documentation: Metadata

In this step, you will examine a portion of the metadata that gets created automatically at the feature class level.

☐ Open the Catalog window and dock and pin it in the open position.

☐ Navigate to the soil_analysis geodatabase.

☐ Right-click the StudyArea_UTM feature class and choose Item Description.

☐ In the Item Description window, scroll down to the ArcGIS Metadata and expand it.

This is not yet complete metadata—important descriptive information is missing; however, the software automatically populates the metadata document with valuable information.

☐ Within the ArcGIS Metadata section, expand ESRI Geoprocessing History.

☐ What four tools were used to create the study area boundary?

1. ________________________________
2. ________________________________
3. ________________________________
4. ________________________________