Chapter 03: Finding Data for ArcGIS

Michael W. Pesses, Antelope Valley College

Introduction

Good GIS starts with good data. Your analysis and resulting maps will be useless if you don't have the proper inputs. Finding good data can be a major challenge and this chapter will show you what to look for and where to find it.

When you first open ArcMap, you are faced with a haunting white space and dozens of cryptic buttons. You need to add data of some sort to begin to fill that white space and begin to use the many tools that are at your disposal. An analogy is the program *iTunes*. In and of itself it is somewhat useless. You need to add *mp3* files or stream data to actually play music and use the program. This also ties in with the need for good data. You can add music to iTunes to play at a party, but if you add terrible music all of your friends will leave. The Dave Matthews Band is a good example. No one wants to hear that. Now if you play some *Exile on Main Street* you've got a party. Same thing with GIS data. Dave Matthews quality data will produce bad results; the Stones will give you the answers you want.

This chapter will link to some sources for GIS data, but they may have moved, altered, or taken down since this chapter was last updated. First, feel free to contact me (mpesses@avc.edu) if you find that this is the case. Additionally, rather than click on the following links, a Google search for the data you need can sometimes be easier. For example, if I am working on a project about grey wolves in North America, I would enter "grey wolf GIS data" into the search bar and see what I can find.

Data can be stored in a variety of formats that work with GIS. Knowing what file types work well with ArcGIS is important, as well as how these files are stored in the Windows operating system. This list is not complete, but you will be introduced to the most common forms you should expect to work with.

GIS Data Basics

Before I get into specific file formats, it is important to think about the nature of GIS data. First, GIS data are spatial in nature, meaning they are connected to a location of some type. This location can be specific point like an address or set of coordinates or can be broader like a country or continent. We can also model these data in one of two ways. **Vector data** are those that are **discrete**, meaning they have a beginning and end. We can use points, lines, and polygons (an enclosed shape) to represent these in the

GIS. The state of California, for example, has clearly defined boundaries, which means that using vectors would be the best way to model those boundaries. In GIS, we often refer to vector data as **features**. A vector data file, like a data file containing the boundaries of all 50 states in the United States of America, is called a **feature class**. **Continuous data**, however, are those data which do not have clear boundaries. For example, elevation data exist everywhere on the planet. There is no place that does not have some value for its elevation. Coastal places might have an elevation of 0' above sea level, but this number is still elevation data. We use the **raster data** model for these continuous data. A raster is a grid of cells, or pixels, which are assigned a value. Because raster data are based on a grid, resolution becomes important. A raster with a resolution of 30 meters means that each grid cell covers an area equivalent to a 30 meter by 30 meter square. Clearly, you want to have a raster with the best possible resolution, but the finer the resolution the harder it is to collect the data and the larger the size of the file. A balance between resolution and having a data file that your computer's processor can handle is necessary.

Shapefiles

The simplest format for vector data is the shapefile, which can be used by ArcGIS as well as other programs like the open-source QGIS. Shapefiles are simple, but that is not necessarily a bad thing. While ESRI pushes for the use of their proprietary geodatabase format (discussed below), shapefiles are an easily sharable way of storing spatial data that are almost universal within geospatial technology. Shapefiles can only store one type of feature class, meaning an individual file will hold either point, line, or polygon data. The file itself is actually comprised of at least three individual files that a program like ArcGIS will read as one file. Figure 03.1 below shows how the Windows operating system "sees" a shapefile for major roads in the United States. Note that the seven individual files all have the same name, "mjr_rds," but different extensions like .shp and .dbf. Shapefiles must have a .shp file, which contains the drawn geographic data, a .shx file which the computer uses for indexing the geometry, and a .dbf file which contains the attribute table for the data. Additional files contain projection information, additional indices, and metadata. It is very important to remember that if you want to move your shapefile to another location on your computer or the cloud, or you want to share it with a colleague, you must keep every file together or it will no longer work. The same is true of renaming a shapefile. If you rename it through Windows' File Explorer, you must rename each separate file with the exact same name.

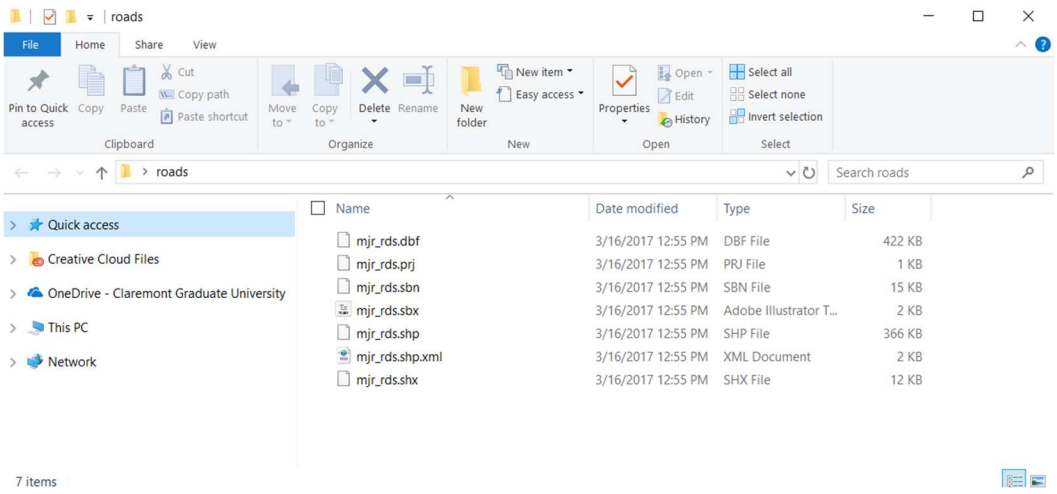


Figure 03.01. How the Windows operating system sees a shapefile.

ArcGIS is programmed to recognize that each file with the same name is part of the same shapefile. Figure 03.2 shows how ArcGIS will display the shapefile. All of the separate files have been merged into a single "mjr_rds.shp" file. The other files still exist, but ArcGIS hides them from us to make everything simpler. If you rename the shapefile through ArcGIS, it will individually rename each separate file. Also note the green icon next to the shapefile. The green color tells you it is a shapefile and the line design indicates that it is a line feature class. Point and polygon shapefiles will also have a green icon but will have a different design to indicate the type of feature class.

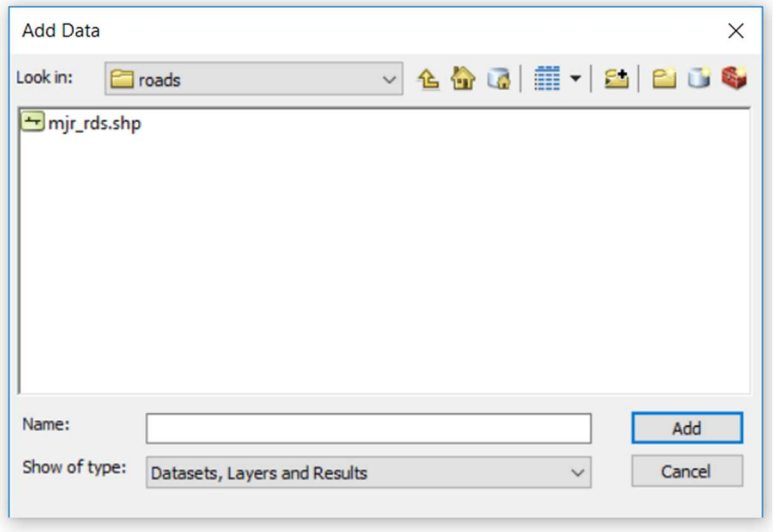


Figure 03.02. How ArcGIS sees a shapefile.

When a shapefile is added into ArcMap, the program uses the .shp and .shx files to draw the actual map data. Again, this is a linear feature class, so the only data stored in the shapefile are lines representing major roads in California. This particular shapefile also has a .prj file, which contains the necessary projection and coordinate system information so that ArcMap can place it properly in space.

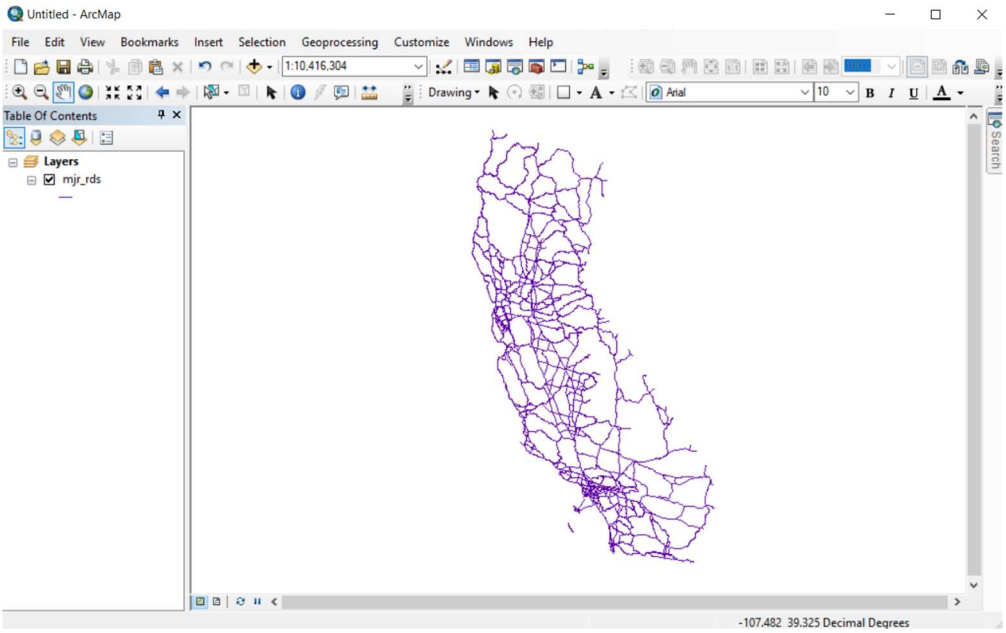


Figure 03.03. Adding a shapefile to ArcMap.

Right-clicking on the "mjr_rds" layer in the Table of Contents window on the left of the screen and selecting "Open Attribute Table" will display the contents of the .dbf file (see Figure 03.4). Each line drawn in ArcMap is connected to a row in the attribute table which contains information. In this case, each line has information like the name of the road and the type of road it is. This attribute table information can be used for generating labels or conducting spatial analysis regarding specific roads.

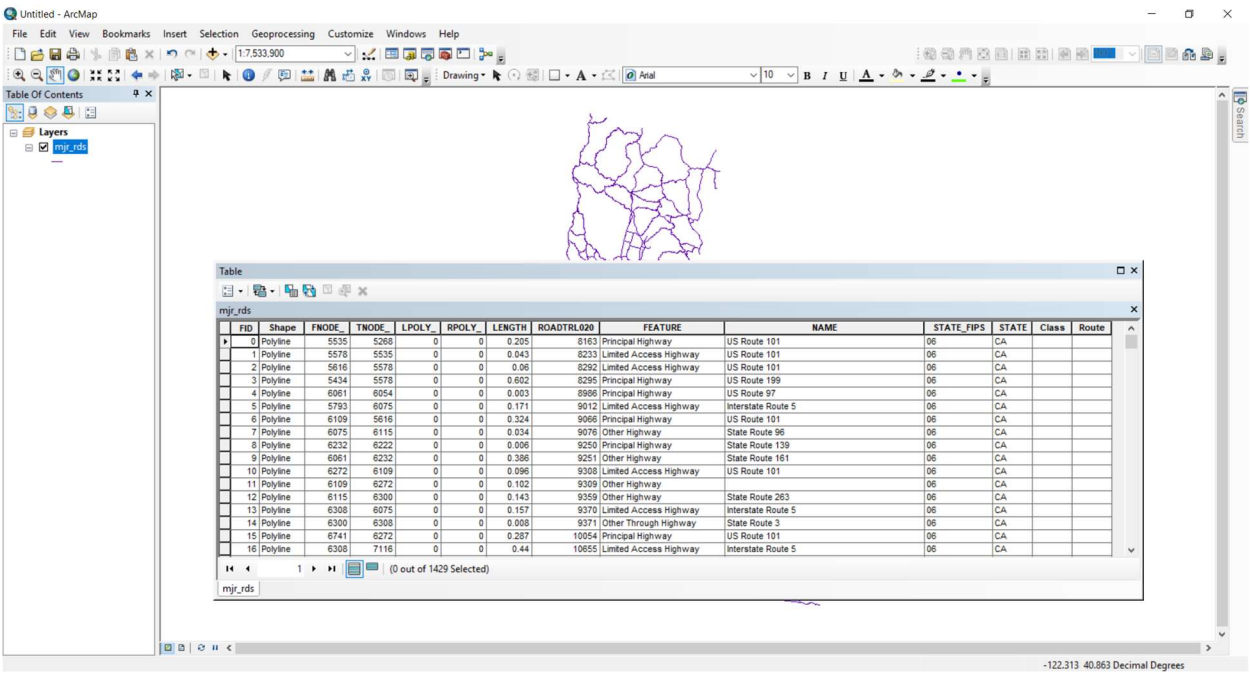


Figure 03.04. Displaying a shapefile's attribute table.

Finally, you should note that the name of the shapefile does not contain spaces or special characters. When naming any type of data for use in GIS, always avoid spaces or overly complex names. Using an underscore "_" is a safer way to break up words in a file name. This is not quite as important as it once was with earlier versions of ArcGIS and its predecessors ArcView and ARC/INFO. A single space in one file could cause the entire program to crash. While the newer versions of the software are much better, ArcGIS still can get hung up on little things like a file name. Instead of calling the file "Major Roads in California.shp" I stuck with the simpler and shorter "mjr_rds.shp" to avoid trouble when I start analyzing and manipulating the file in ArcGIS.

Geodatabases

ArcGIS build their proprietary geodatabase formats to store multiple feature classes in one place and to increase how we can make connections across those feature classes. ESRI has developed several types of geodatabases over the years. An **ArcSDE Geodatabase** is for "enterprise" purposes, meaning it is to be shared over a large organization with multiple users. For more localized work on your own computer, ESRI also developed the **Personal Geodatabase** which was built off of the Microsoft Access database format. These were limited in size and capability and have all but been replaced with the newer **File Geodatabase** format. Regardless of type, a geodatabase is

a relational database, meaning all data are related through a series of tables that ensure unique values are kept independent and are easily maintained. The end result is an efficient way to store data. Figure 03.5 shows how a group of shapefiles can be stored in a file geodatabase.

Figure 03.05. Shapefiles and raster images brought into a File Geodatabase.

The silver cylinder at the top is the entire geodatabase itself. Underneath it are three squares and the name "georeference." This is a **feature dataset** which can contain one or more feature classes of any vector type. A feature dataset is a great way to group like data. Further, a dataset can be set to a specific coordinate system, which ensures that all features within it are consistently placed in relation to one another. This becomes important when you start doing geoprocessing and spatial analysis. Also note the "land_use_Topology" file in the topology feature dataset. This is not a feature class, but actually a set of rules that can be enforced to ensure that data behave properly in relation to one another. For example, in this land use example, the general plan data should not overlap with one another, meaning that a place cannot be zoned as both residential and commercial. The topology rules will enforce this fact to ensure that anyone editing the data does not accidentally make this mistake.

Raster Images

Raster data can come in a variety of file formats. To represent continuous data, rasters are made up of grids of cells, each of which store data. We use the term **pixel** to describe these cells in digital photographs. A digital photo is comprised of grid cells that represent a single color. When placed together, all of these cells make an image.

Esri's Grid format is a classic way to store raster data in ArcGIS. These are useful in the GIS environment, but cannot always be used in other programs. Figures 03.06 through 03.09 show a grid file that was downloaded from http://www.soest.hawaii.edu/coasts/data/hawaii/dem.html.

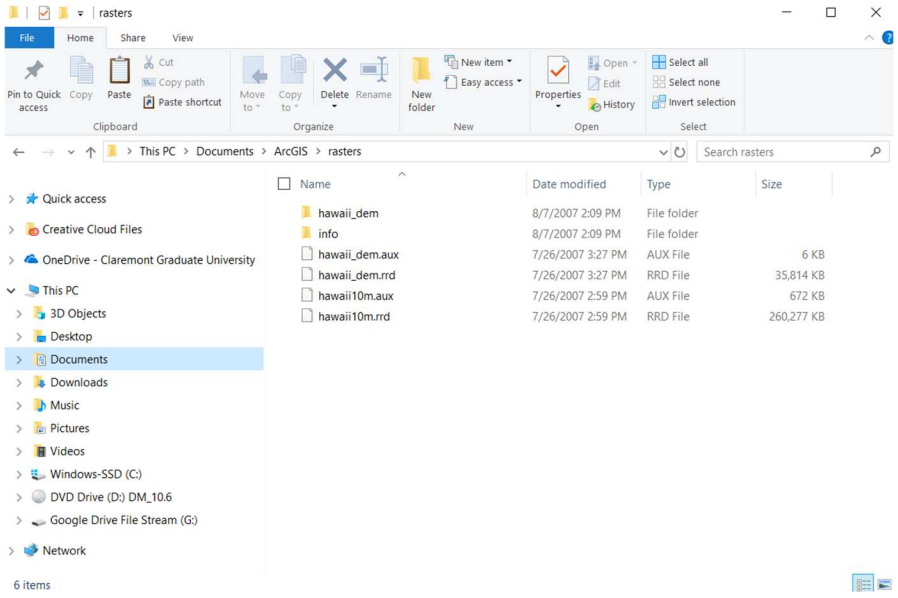


Figure 03.06. A digital elevation model (DEM) of the big island of Hawai'i stored as a Grid file. The folders contain key files that tell ArcMap how and where to display the raster.

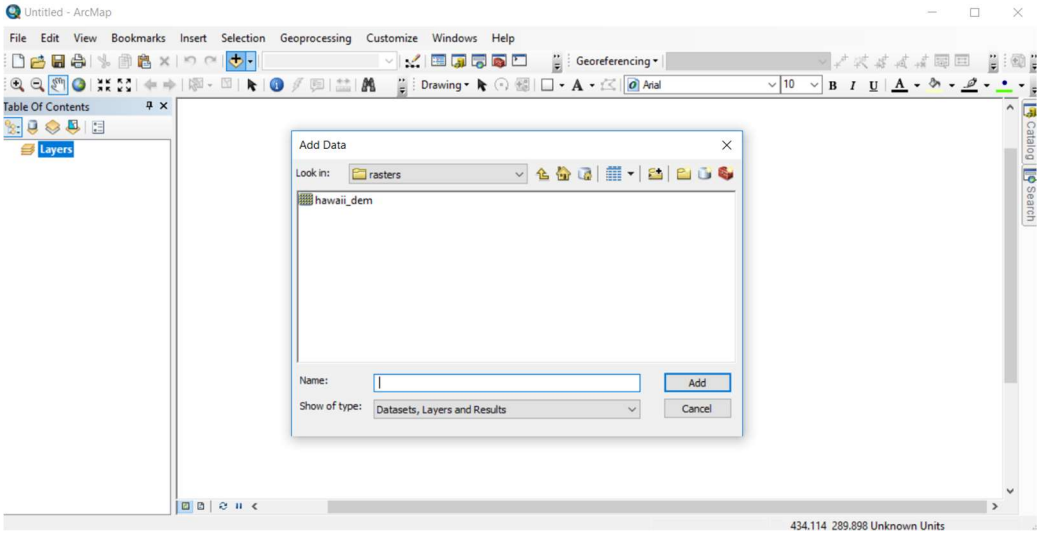


Figure 03.07. ArcMap shows the DEM as one file for simplicity.

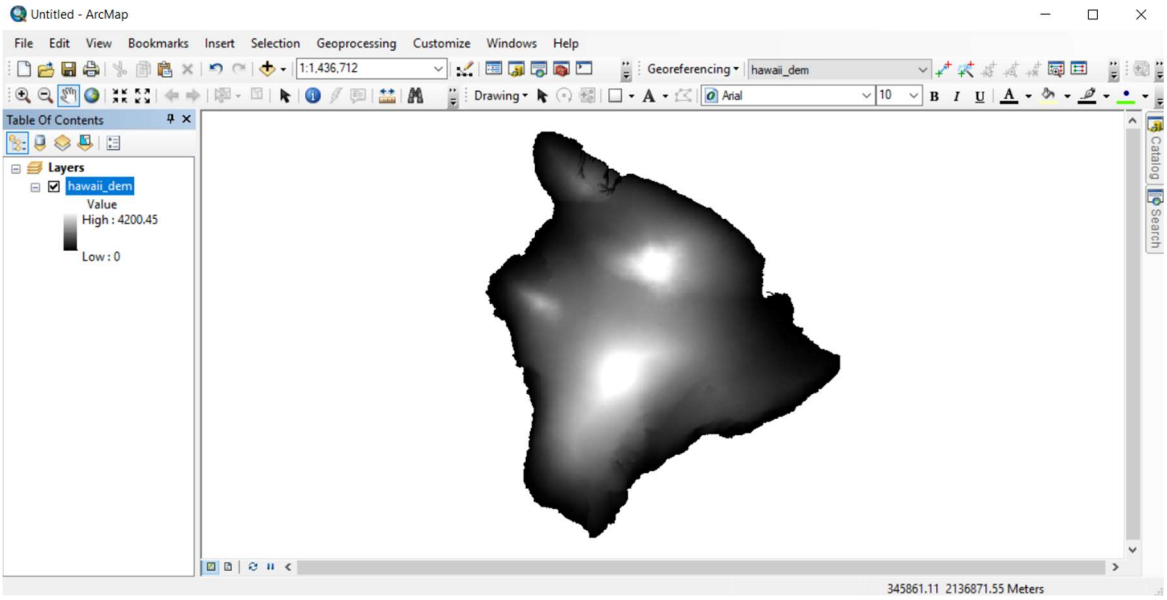


Figure 03.08. The DEM represents a single value of elevation within each cell. It is displayed here ranging from black to white, with darker colors representing lower values.

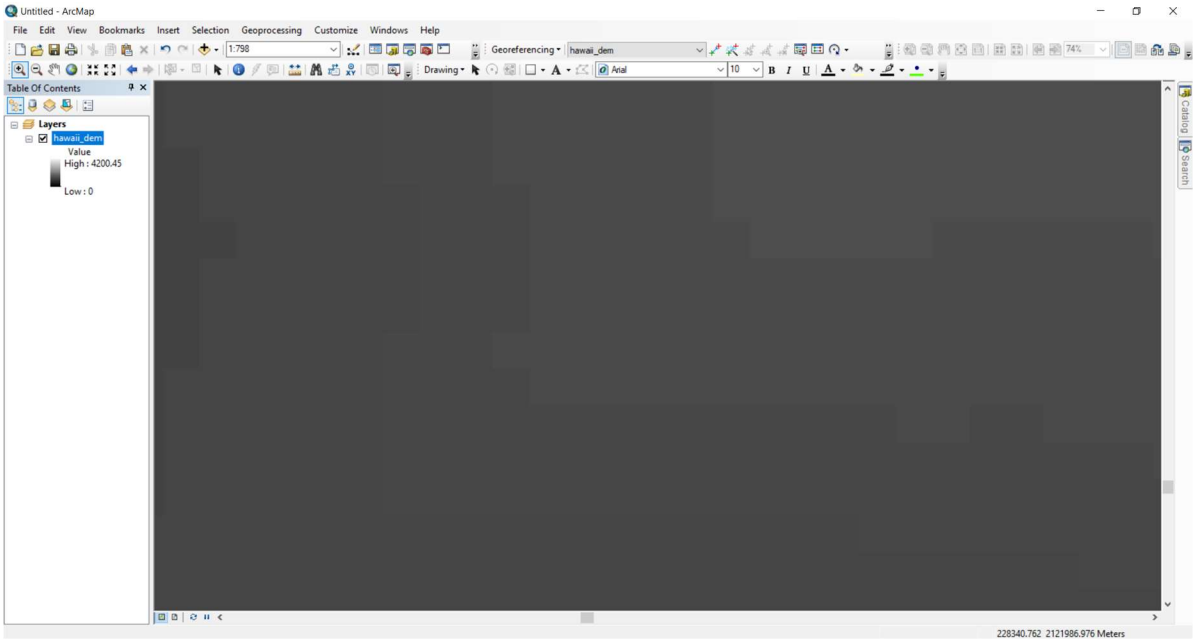


Figure 03.09. Zooming into the raster shows the individual cells, with are at a 10-meter resolution.

ArcGIS can handle many other types of raster data (a full list can be found at http://desktop.arcgis.com/en/arcmap/10.3/manage-data/raster-and-images/supported-

raster-dataset-file-formats.htm). Tagged Image File Format (TIFF) and the related GEOTIFF format is useful for storing scanned maps that can be placed in the GIS at its actual coordinates. Scanning a map and placing it in space at its actual coordinates is a convenient way to digitize older analog data. The **georeferencing** process, which is the act of placing the file in space, is explained in Chapter 08.

Tables

A useful source of data are tables which are connected to space in some way. By being connected to space, I mean that the table either contains coordinates of some kind for point data or is has broader information like state or county names. For example, we could track down a table of the percentage of people living below the poverty line. If the data are separated by state, we can then connect it to a feature class of the United States. We can connect a table to a feature class by performing either a **join**, in which the table is attached to the feature class's attribute table, or a **relate**, in which the attribute table and new table are linked. To perform either process, the new table and the feature class's attribute table must each have a field that matches. To use the poverty line example, you could join the two tables if they each had a field that contained the names of the states. ArcGIS knows to attach the poverty line data for Alabama to the mapped polygon of Alabama because they have the same name. It is important to remember that the two fields must be identical. If the feature class has 'Alabama' but the table has 'ALABAMA' or 'AL' the join will not work. Always remember that ArcGIS works using very simple logic and cannot anticipate or assume our intentions.

Chapter 06 explains tabular data and joining in the context of using color to represent quantitative values.

Finding Data

The internet has a tremendous amount of data, though a good portion of that is garbage. Anyone who has spent just one hour surfing the web is well aware of this. The same goes for online GIS data. Be careful when you are searching. For example, unless you are working on a professional project, ***you should never pay for data***. Most GIS data is produced by government agencies, and therefore is public and should be offered freely. Also, some data may be good, but are at the wrong scale for your intended purpose. For instance, if you are making a map for fieldwork, i.e. one a person will take to explore a riparian study area, you should not use stream data that are used for world

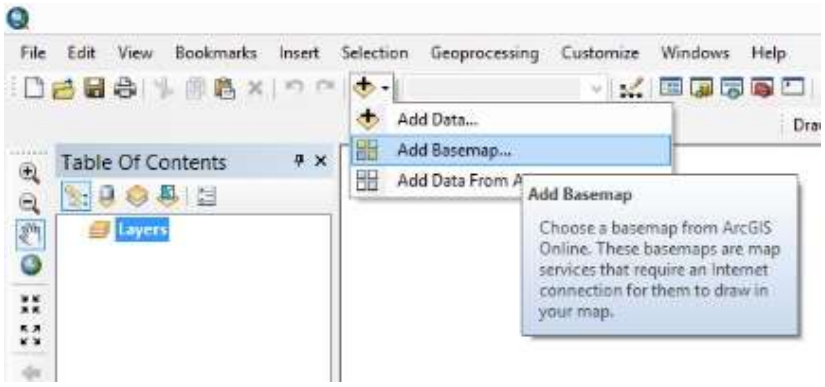
maps. The small scale of the data means that the detail is not enough to be of any use in the field. Reference the section on scale in Chapter 02 for more information.

The following sections show some good sources for GIS data, though they might not be the best data for what you have in mind. Ask these three questions when you find a dataset online:

1. **Who created this dataset?**
2. **Is it the right scale for my purposes?**
3. **Is it up to date for what I am studying? For example, does it show population densities from 1990, but nothing for the 21st Century?**

ArcGIS Online

You can connect to the ESRI servers and download basemaps, a term we use for the underlying data on a map. This can be shaded relief, aerial photos, scanned topographic maps; a basemap basically gives you the context and location for whatever it is you are mapping. For example, if I am mapping crimes in Los Angeles, I could add a premade basemap to show the streets and landmarks.



To access the ArcGIS Online Basemaps, simply click the dropdown next to the yellow "Add Data" button and select "Add Basemap." You can select any of the available maps and it will be added to your map. While these maps are a nice feature of ArcGIS, they are quite limited. You cannot change them or use them for analysis. It also looks much more impressive if you can make your own basemaps rather than rely on ESRI's.

World Bank

http://data.worldbank.org

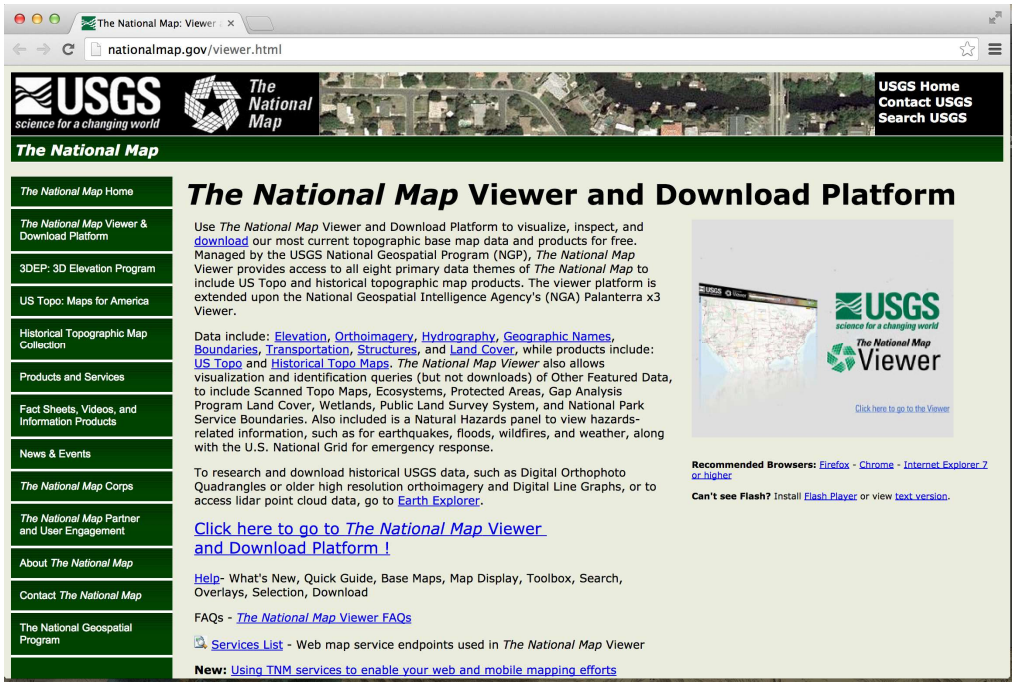
The World Bank funds development projects around the world, and provides a wealth of data on demographic, health, economic, social justice, and environmental issues. The data are accessible in tabular form. Bringing the data into ArcGIS will be addressed in Chapter 5, but there is a map viewer that will allow you to preview the data.

The National Map

nationalmap.gov

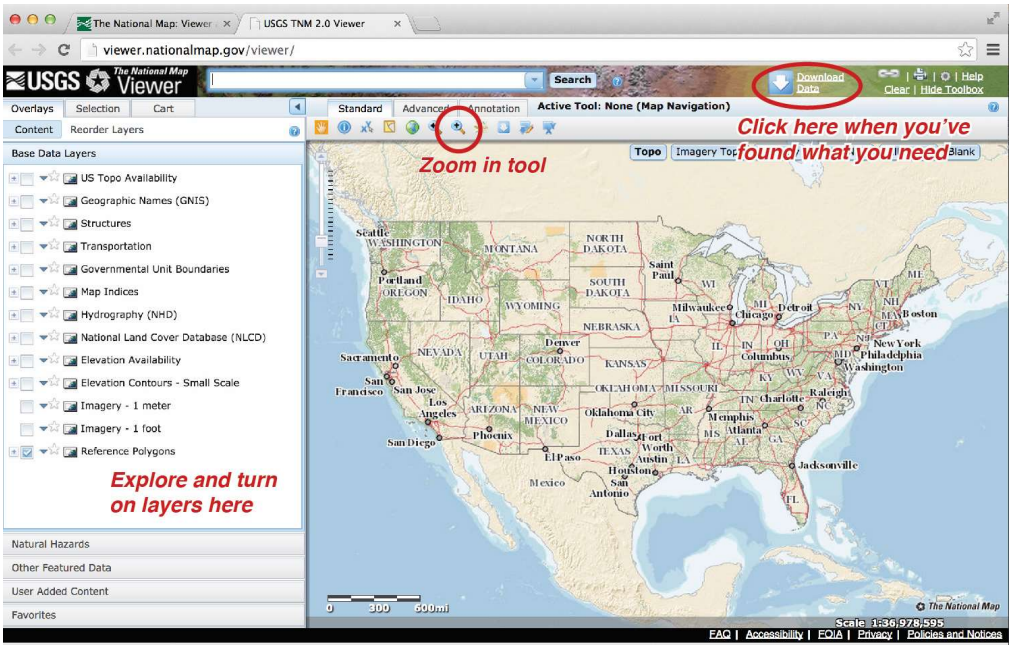
The National Map is a portal for all of the United States Geological Survey's (USGS) mapping projects. In an effort to streamline resources, the USGS is merging previous data sites into this one. For example, the National Atlas dataset is being merged over. It is a series of 1:1,000,000 data that can be useful for mapping things across the country, and maybe even in California, but not in smaller states or for cities.

There is a lot of information on this website, but the main thing you will want is the *National Map Viewer*, described on the site as a resource for "more experienced map makers and professional geographic information users." If you have read this far, you can consider yourself "more experienced" than the average person, so click the link!



National Map website

This site will allow you to download elevation data, imagery, boundary hydrography, and some transportation data. The download process can be a tad cumbersome, but easy once you get the hang of it. The link to download the data will actually be emailed to you during the process. Once you click on that link in the email the download will begin.

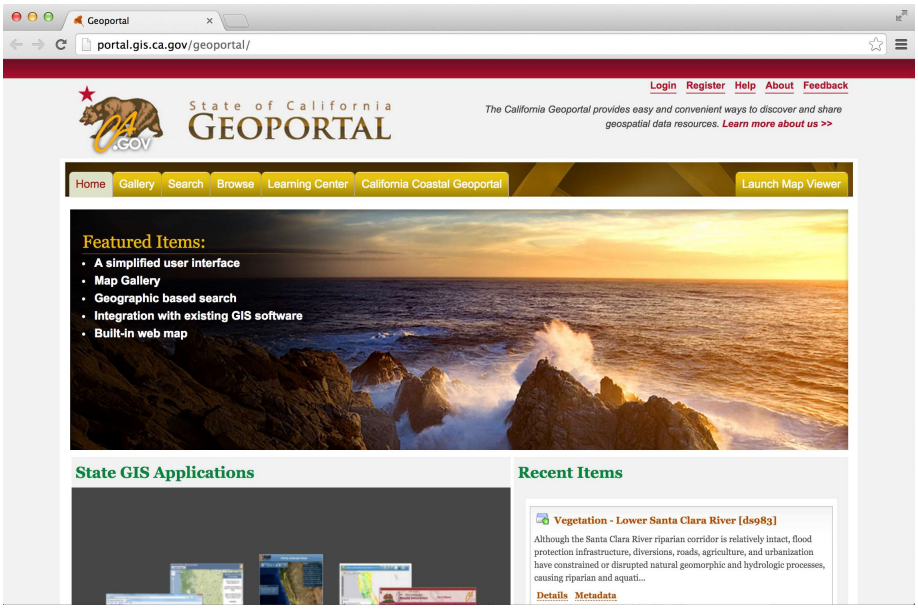


The National Map Viewer, from where you actually download the data

State of California Geoportal

http://portal.gis.ca.gov/

California has collected a good deal of data for the state in their "Geoportal." In an effort to make it easier to use, I think the state has made it more confusing, but you young kids may be okay with it. Before the downloadable data were available in list form, but now the state uses a map viewer similar to the USGS and a "searchable" browse feature.



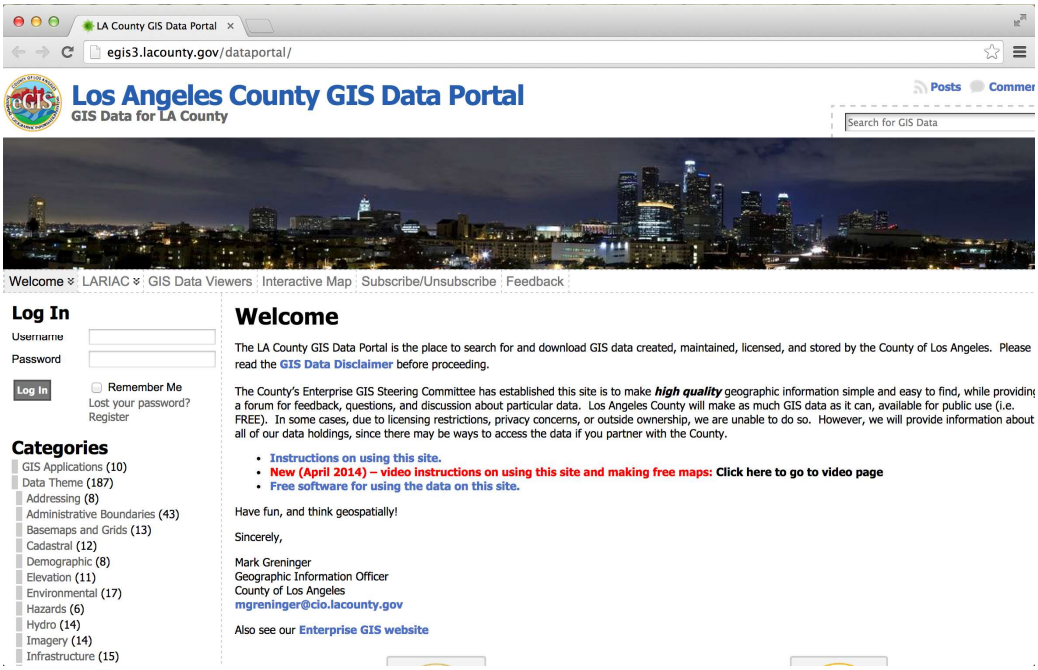
State of California's Geoportal

Ignore the map viewer. It doesn't give us much. The Browse tab is the most useful for the "more experienced map makers and professional geographic information users" as the USGS puts it. Click on it and then select the downloadable data option. You can scroll through the pages of data, though unfortunately they aren't in a friendly order. You can add a word or term to "filter" the results, but unless you know the exact name of a file this can be more difficult than sorting through the unfiltered results. When you get what you want, click on the "open" link and it will begin to download a compressed file of the dataset.

Los Angeles County GIS

http://egis3.lacounty.gov/dataportal/

Los Angeles County also maintains a portal for the various data sets to try to have one space for the products of multiple county and city governments in the region. You can use the "Interactive Map" to zoom in and visualize if you must. This site is designed for easier browsing though. Simply, click on the Categories/Data Theme and explore the links to find something that works for you.



US Census Bureau TIGER Files

http://www.census.gov/geo/maps-data/data/tiger.html

TIGER stands for Topologically Integrated Geographic Encoding and Referencing, which sounds like someone just wanted to use the word tiger and forced an acronym. The Census Bureau is known for counting the number of people in the country, but

they also maintain a database of roads, landmarks, as well as demographic units called blocks and tracts.

Unzipping data

File compression, or "zipping" as it is sometimes called, is a way to shrink the size of a file or files to be able to easily download or email it. If you received a compressed file, you will need to "regrow" or "unzip" the data to be able to use it.

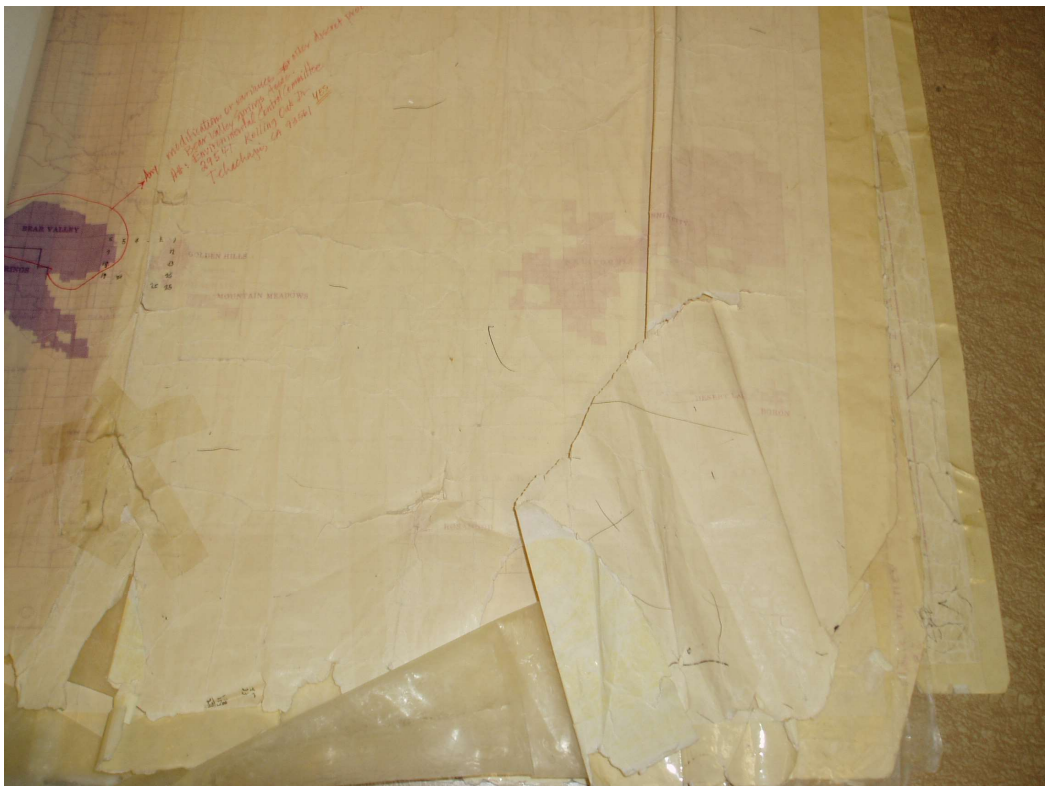
Windows computers do come with preloaded compression software that works on "zipped" files. You will know you're working with such a file because the file extension is ".zip" (e.g. gis_data.zip). Simply right click on the data file and "unzip" or "extract" the file to be able to use it.

Some websites will offer data that are "double compressed," often with a ".tar.gz" extension (e.g. gis_data.tar.gz). As of this writing, Windows can't handle these. You will need to download a program like 7-zip, an open source "file archiver" which can be downloaded from http://www.7-zip.org. Once installed, right click on the downloaded file. You will now see an option for 7-Zip. Select "extract here" or "extract files." This will get rid of the ".gz," but you'll get a new file that simply has the ".tar" extension. Repeat the process one more time to be able to access the data.

A Cautionary Tale

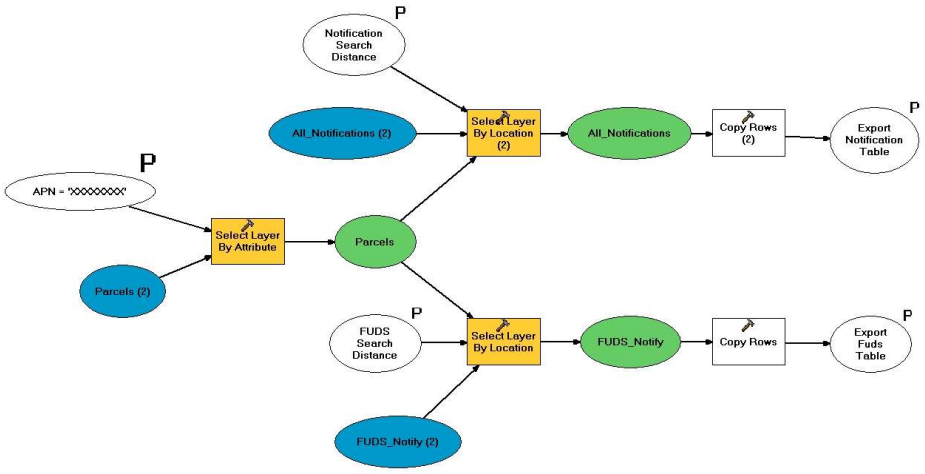
I was hired by a planning department to introduce GIS and bring their operations into the 21st century. A planning department handles many things regarding land use within an area, and one of the jobs planners did was to map out a 1,000 foot radius around a project and alert other agencies and property owners. Before GIS, planners would walk to a table that held a stack of about forty maps, locate their project's area on each one, mentally draw out a 1,000 foot circle, and then write out each agency or person that fell within that circle. This seemed like an incredible waste of time to me, and a perfect way to use GIS to streamline work in the office. It was not an easy task however, as you can see from the pictures below. No one really knew just how old the maps were; people just updated them when need be, often using things like thick-tipped highlighters to do so. The maps were also heavily ripped and some of the information had faded so much that the information was lost forever.

One of the infamous "Notify Maps." Note the thick highlighter line and yes, questionable stains.



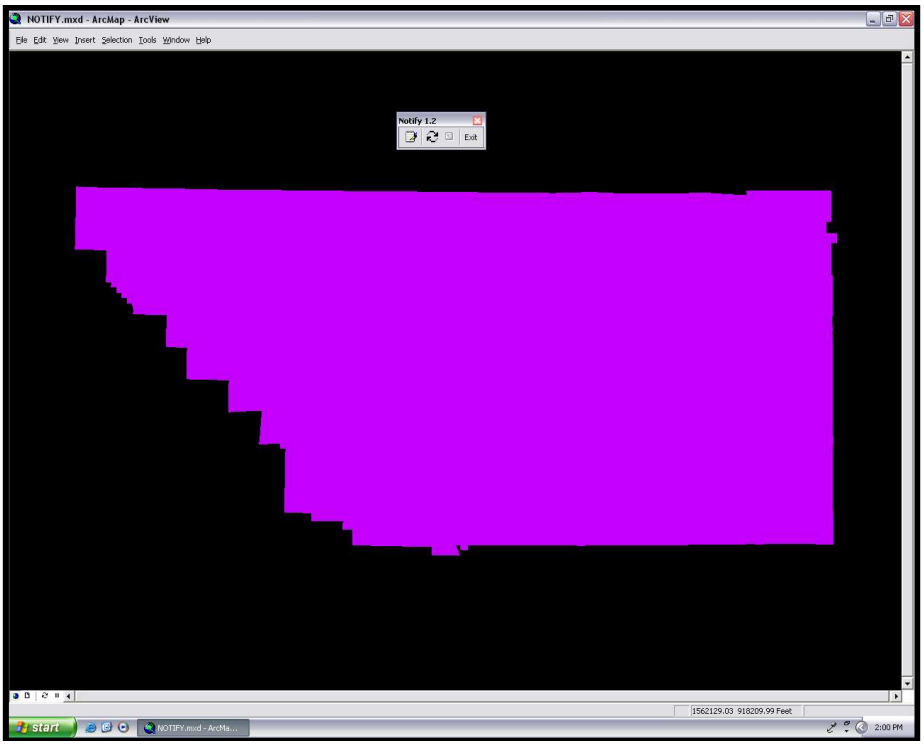
Another "Notify Map." Too ripped to send through the scanner, and so faded that a lot of information was lost.

I scanned as many of these as I could and digitized the data into a geodatabase, with the idea that planners could type in their project and have the computer do the work. Once the data were digitized, I constructed a workflow for the computer and turned it into a simple button for planning staff to press and then enter in some relevant information. The output was a Microsoft Excel table with the names and addresses of people needing to be notified. This table could then be used to automatically generate address labels.



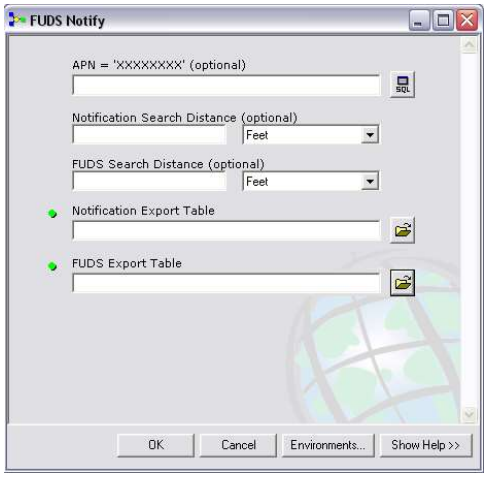
The workflow for the notification program as outlined in ESRI's Model Builder.

So when all of this was done, the planners (many of whom were not "more experienced map makers and professional geographic information users") got a very simplified interface with which to work.



What the planners saw

They pressed a button and up popped a window that requested certain information. The whole process now took two minutes.



This window popped up when the planners clicked the "notify" button. APN was for their project's property. FUDS referred to a separate Federal project that we later incorporated with the notification process.

Now here's the word of caution. The new interface was clean and looked very official, but the data were still the same! While the planners could now save time, it was important not to lose sight of the fact that the data was pretty poorly implemented and maintained before this. For example, the highlighter lines on the maps were actually

one half mile thick on these maps due to the scale. This took some work on my part to figure out the intent of these lines. The digital data therefore were only as good as me.

So just because the data are digital and come from a government office does not mean that they are infallible. Question the data you find to ensure that your work is not doomed from the beginning!