Understanding Geospatial Data Files used in Geographic Information System Software

A Basic Introduction

by
Lynn Carlson
GIS Manager
Environmental and Remote Technologies Lab
Brown University

Understanding Geospatial Data Files

A: What Is A Spatial Data File?

1) Spatial data files are somewhat like other files you work with on a computer. They can be:

a) stored on a hard drive, memory stick, CD, DVD
b) assigned either a user-defined file name, or are given default file name by a software application
c) organized into folders
d) have the ability to be opened, viewed and edited by one or more GIS software applications that understand their format.

2) However, that is where the similarities end. Spatial data files are unique in that they store "georeferenced" information – information that defines location or place. In addition, descriptive information about the georeferenced information is stored in each spatial data file.

3) Thus rather than just text (like a word processing document) or numbers (like a spreadsheet), an individual spatial data file is a digital representation of a similar group of geographic features on the surface of the earth (or any other planetary body!).

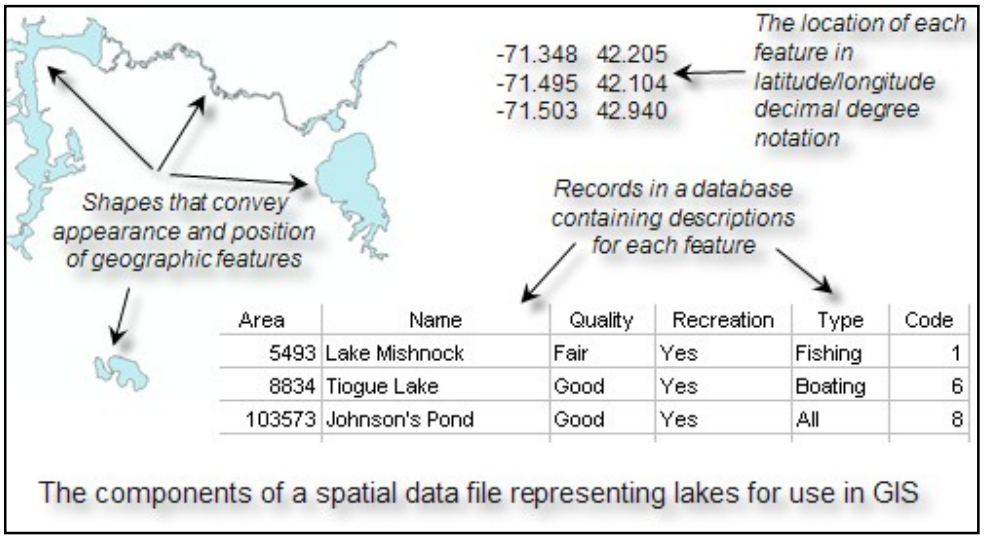
4) The geographic features can be actual physical entities or events, or they can represent conceptual features.

5) Examples of individual spatial data files representing real geographic features or events are lakes, rivers, wetlands, elevation contours, roads, forested areas, rare species habitats, soils, earthquakes, vehicle thefts, electricity distribution lines, and groundwater reservoirs.

6) Examples of individual spatial data files representing conceptual geographic features are census tract boundaries, zoning boundaries, or parcel boundaries (i.e. conceptual features do not physically exist on the landscape, but are imposed by us for various reasons and can be represented in a geographic context).

7) Each spatial data file is uniquely constructed to work within GIS software applications. Each one consists of unique characteristics:

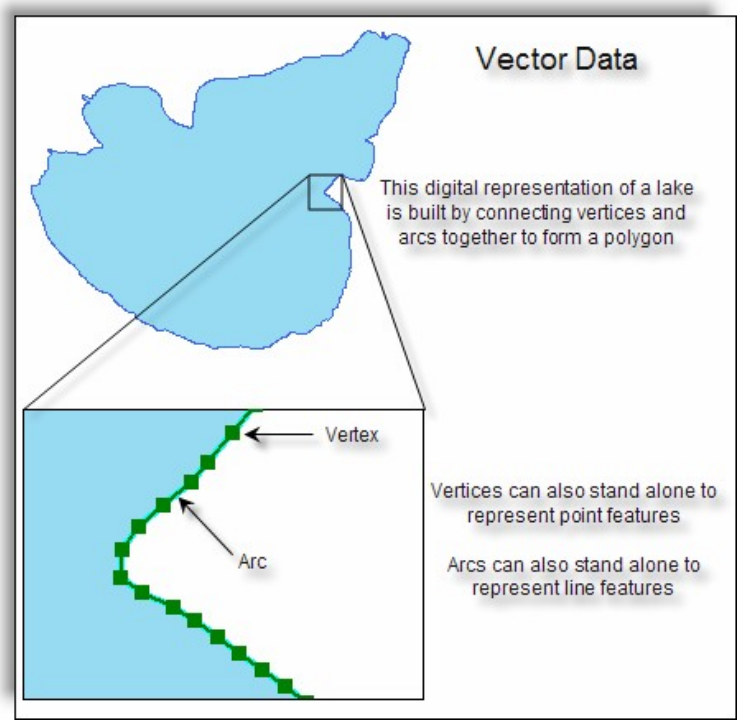
a) "shapes" that attempt to reflect / convey the appearance and position of individual geographic features as accurately as possible
b) records within a related tabular database that contain numeric and/or textual descriptions of each feature
c) a coordinate system that defines the true location of all the features on the earth's surface (i.e. the latitude/longitude)



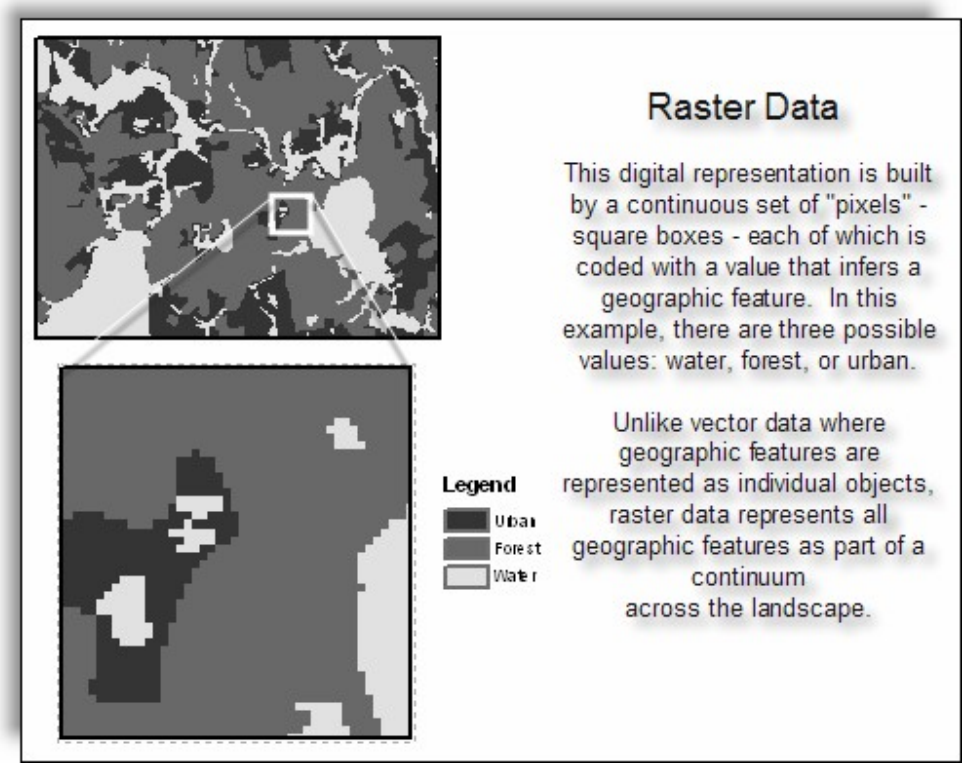
B: Formats of Spatial Data Files

Spatial data files come in several different formats. You may need to use only one, or you may need to use a combination of them, depending on your particular application and/or type of analysis. Each format falls under one of two different categories: **vector** or **raster**.

Vector spatial data files are ones in which the geographic features being represented are built by a collection of vertices and lines.



Raster spatial data files are ones in which the geographic features across an entire area are represented by a continuous set of "pixels" or "cells".



1) The <u>Shapefile</u> spatial data file format

a) This is a very common format for spatial data files in the **vector** category.

b) In this format, geographic features can be represented in one of three ways:

i) points
ii) lines (aka arcs)
iii) polygons (aka areas, polylines)

c) If you utilize an existing shapefile, the point, line or polygon representation was chosen by the individual / organization responsible for its development. If you create your own shapefile, you will have to determine which representation is best for your work or application. The determination is made based on several factors, including but not limited to:

i) the need to depict features at a specific scale (e.g. cities across the U.S. as points vs. cities within a county as polygons);

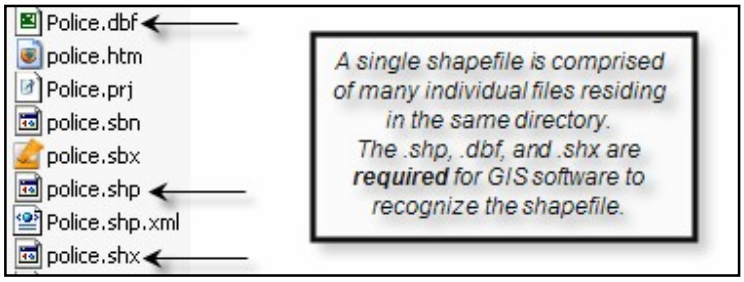
ii) the need to depict features as realistically as possible (e.g. the center line of rivers vs. the width of the entire river with both banks constituting a polygon);

iii) the need to quantify some aspect of geographic features (e.g. the size of a lake would require it be created as a polygon vs. the length of the lake's shoreline requiring it to be created as a line).

d) A single shapefile is actually comprised of three files that absolutely **must** reside together in the same directory on your disk, or else it will not be recognized by the software. These three files will always have the *same* prefix for their filename, but have *different* extensions.

i) the **.shp** file contains the shapes (e.g. rivers.shp)
ii) the **.dbf** file contains the tabular database records (e.g. rivers.dbf)
iii) the .**shx** file contains an index which connects the table to the shapes (e.g. rivers.shx)

e) While not required for GIS software to recognize a shapefile, a **.prj** file (e.g. rivers.prj) is very important. This file contains the *coordinate system* definition for the shapefile. You will work more with coordinate systems and .prj files in a later chapter. For now, just recognize that a .prj file, while not essential, is very important.



f) After you have worked with a shapefile, you may find that additional files appear in the directory along with the three "base" files. For example, **.sbn** and **.sbx** are common. Each one of these "extra" files is generated by the GIS software for a specific purpose (e.g. to speed drawing time). Even though they are not required for the software to recognize the shapefile, it is generally preferable to retain them unless you have a specific reason or need to delete them.

f) If you obtain a shapefile from another organization, person, or from a website, it will often be provided as a compressed .zip file. Once you have unzipped the file (using a utility program such as WinZip, PKzip, or QuickZip), check to see that, at a minimum, the three base files are in the directory. Otherwise, the file will be useless.

If you do not know how to use a utility program to unzip files, please ask for help!

g) While still very common, the shapefile format is slowly being superseded by a new format - the *geodatabase* (see below).

2. The <u>Coverage</u> spatial data file format

a) This was the original spatial data file format used in GIS software. While this format has taken a "backseat" to the shapefile format due to the simplicity of shapefiles, coverages are still very viable and have many advantages.

b) Along with shapefiles, the coverage format is being superseded by the new geodatabase format (see below).

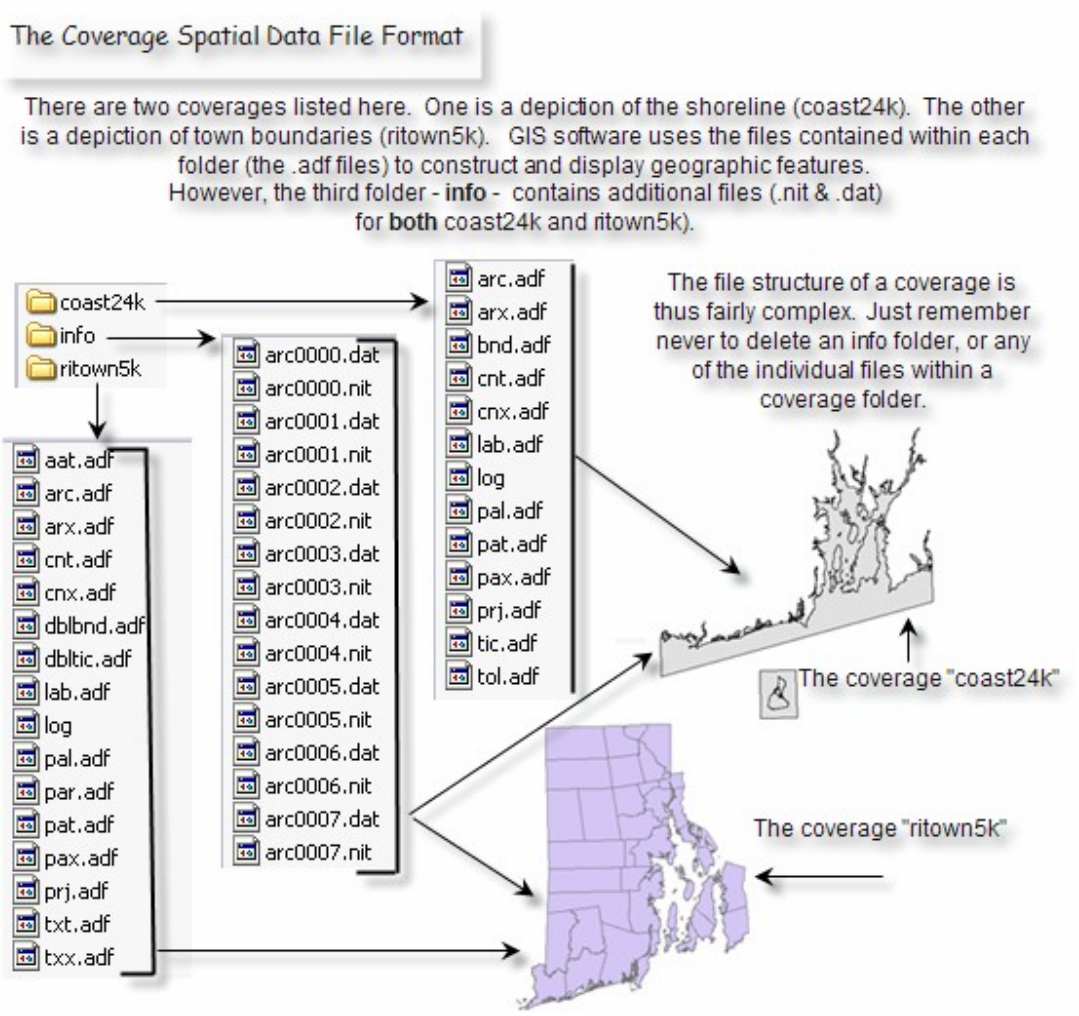
c) However, many web sites still offer spatial data for download in the coverage format, so you should at least know that they exist, and know a little bit about their structure in the event you need to use one.

d) Just like shapefiles, geographic features are represented as points, lines, or polygons and many factors come into play when deciding which representation is best (see B-1-c above). Coverages also fall within the **vector** category.

e) Unlike shapefiles, a single coverage is actually comprised of two ***folders***. Each folder contains a multitude of other files that the GIS software "puts together" in order to represent geographic features and associated tabular information when it is opened.

f) If either folder is missing, or if files from within either folder are missing, the coverage will be "corrupt" and not useable.

g) The following graphic is meant to aid your understanding of how coverages appear on your hard drive:

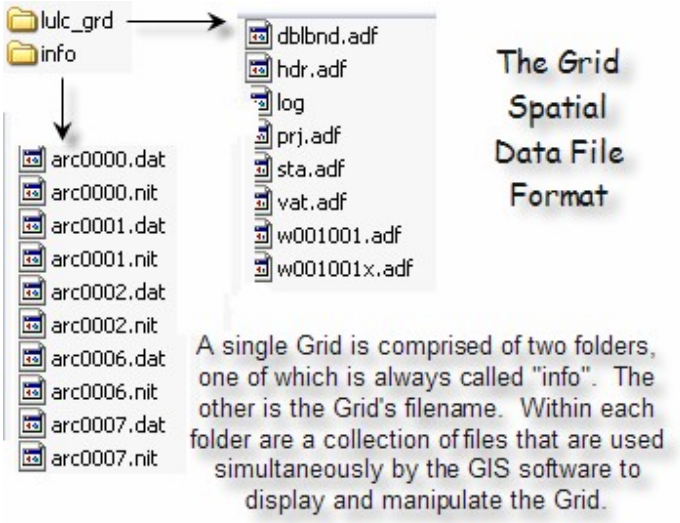


h) Coverages and shapefiles are often used almost interchangeably in GIS. They each can represent the same geographic features. It is only the internal file structure that is different. An analogy would be a Microsoft Word document vs. a Corel Word Perfect document. Both files are used to contain text (primarily) and you can import and export them at will, but they have different underlying structures which are, most of the time, invisible to you. Similarly, coverages and shapefiles are both used to contain geographic data of the vector type. It is possible to convert a shapefile to a coverage and vice-versa. Each format has advantages and disadvantages which will be pointed out in later chapters.

3) The Grid spatial data file format

a) In most respects, grids are very different from either shapefiles or coverages. Grids fall into the **raster** category; they are constructed of rows and columns of pixels instead of vertices and arcs.

b) Like coverages however, grids are comprised of two folders, each containing files that the software "puts together" for display and manipulation.



c) Grids can be either:

i) Integer Grids – in this case, the pixel values are integers and each integer may also be associated with one or more textual descriptions.

ii) Floating Point Grids – in this case, the pixel values will be expressed as decimals. Floating point grids can not have textual descriptions.

You will learn more about the distinction of these two types of Grids in a later chapter.

4) Images as Spatial Data Files

a) Many different image formats can be used in GIS. Some of these may be more familiar than others: .jpg, .tif, .bil, .png, .img, .sid

b) All image formats fall within the **raster** category of spatial data.

c) In some cases, images are not used specifically as "spatial data", but are used to enhance spatial data by providing a digital photograph of a place or object. For example,

a shapefile representing all land parcels within a city may have links to digital photographs of each house on each parcel.

d) In other cases, the images themselves are spatial data. Data provided from the Landsat satellite is an example of imagery that is spatial. If you have ever used Google Earth, the images that appear when you zoom in are spatial data.

e) When an image is "georeferenced" - meaning that information is embedded within the image that describes its position on the surface of the earth in real world coordinates (latitude/longitude) – it becomes spatial data.

e) In addition to being "georeferenced", many images may also be "orthorectified". This term refers to a complex process wherein distortions caused by differences in terrain elevation, camera tilt, and edge effects are removed from the image. Images that are both georeferenced and orthorectified are frequently called "orthophotographs" or just "orthos" for short.

5) Computer-Aided Drafting (CAD) files

a) CAD software applications such as AutoCAD and MicroStation produce **vector** data in .dwg or .dxf format.

b) ArcGIS software can read these files directly- they do not need to be imported or changed in any way.

c) In most instances, but not all, these .dwg or .dxf files will include georeferencing information and thus, the data are spatial.

d) For those files that are not georeferenced, there are methods and / or separate software programs that allow them to become georeferenced.

e) CAD files often provide very detailed information such as the floor plan of a building which can be useful for certain GIS applications.

6) The GeoDataBase (aka GDB)

a) The GDB is the newest spatial data file format developed by ESRI (the vendor of the GIS software you will be learning). It is a replacement file format for all of the above.

b) GDBs have very specific advantages over the other file types:

1. All of your spatial data resides within a single file. This eliminates having several different file formats residing in several different folders across your disk.
    - The file has an .mdb extension.
    - Example: a GDB containg all spatial data for the City of Providence might be named providence.mdb
    - Example: a GDB containing all spatial data for Yellowstone National Park might be named Yellowstone.mdb

2. GDBs have greater "intelligence"

    - Geographic features represented in a GDB can be made to "depend" on one another when dependence is needed to more accurately represent de facto relationships.

        - Example: electric power lines are connected across the landscape via towers. In a GDB, spatial data that represents these two separate entities (power lines, towers) can be designed such that, if a tower is moved, all the power lines that are connected to it will also move.

    - Descriptions for geographic features can be set to specific allowable values, eliminating data entry errors. The allowable values are called the "Domain".

        - Example: there are only four possible entries for wetland types in a spatial data file – bog, fen, palustrine, lacustrine. If someone edits the file and tries to code a wetland as "Freshwater", the value will be rejected.

    - Geographic features can have "sub-types"

        - Example: a parcel of land may have an easement encompassing a portion of the parcel. Rather than having two spatial data files, the parcel features within the GDB can be assigned a subtype to accommodate that portion of the parcel that is covered by the easement.

- Annotation can be linked to features.

    - Annotation is a layer of text that appears on the monitor or a map which helps describe some aspect of the geographic features. A good example of an annotation layer is street names. If a street name needs to be changed, this only has to be done once within the spatial data file layer. The annotation layer is updated automatically to reflect the change. Prior to GDBs, you would have to also edit the annotation layer in separate step.

- Spatial data within a GDB can be "linked" to additional tables of descriptive information. This link is written out to a file and is called a "relationship class". You will explore these in a later chapter.

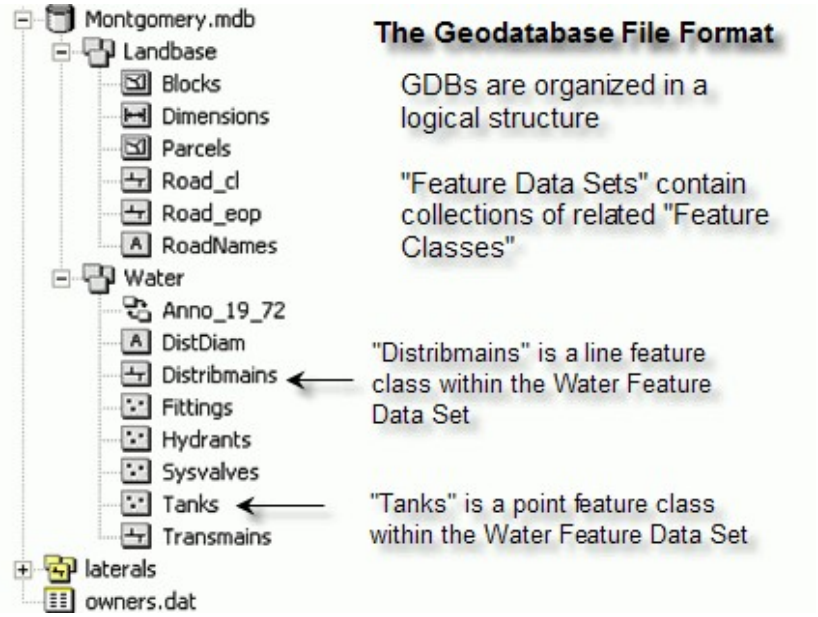
- In a GDB, you can establish "topology rules" between different geographic features to enforce specific conditions when creating a spatial data file.

    - Prior to the existence of GDBs, it was difficult to ensure that geographic features from two different data layers did not overlap in a manner inconsistent with reality.
    - Example: in reality, building footprints should never fall "on top" of lot lines, but when working with two separate data files these type of errors were very frequent.
    - In a GDB, you can literally specify that building footprints are not allowed to overlap lot lines. You can also specify what happens – a warning message or a rejection of data entry – if the error occurs.

3. Industry specific data models can be applied to a GDB

    a) Data models are pre-existing templates provided by the vendor or other organization which contain standardized rules / relationships / data requirements for a GDB so that you don't have to develop your own.

    b) Example: The ArcHydro data model is a template for establishing a GDB of water features on the landscape, along with rules to establish their interaction (e.g. sub-basin delineations must fall within watershed delineations) , a list of data files that are needed (streams, lakes, groundwater aquifers, etc), and a set of tools that are unique to peforming hydrological analyses.



C: Naming Conventions for Spatial Data Files *&* Folders That Contain Spatial Data Files

1) Since spatial data files have a complex structure, it is *extremely* important that you follow some basic rules for assigning names to these files *as well as* to the folders that contain these files.

2) While you will no doubt see other people's files and folders with names that break these rules, I have seen many instances where bad naming conventions cause troublesome results and / or file corruption.

3) Thus, to avoid these problems from ever happening to you, follow these simple rules:

- **Keep The Names Short**
  - Between 8 and 15 characters
  - If you work with spatial data that are in the "coverage" vector format, or the "grid" raster format, only 8 characters are allowed.

- **Do Not Use Spaces**
  - **town boundaries** is a bad name for a file; use **town_boundaries** instead
  - **my gis data** is a bad name for a folder that will contain spatial data files; use **my_gis_data** instead

- **Do Not Use Special Characters (the only exception is the underscore)**
  - e.g. **town#boundaries** is a bad name for a file
  - e.g. **my!gis!data** is a bad name for a folder

- **Never Put A Number As The First Character**
  - e.g. **2000population** is a bad name for a file
  - a good alternative would be **y2000pop**

- **Keep The Names Simple**
  - use abbreviations where possible
  - use underscores where possible as a replacement for spaces
  - use upper case where it makes sense
  - e.g. **twn_bndy** or **TwnBndy**

- **Implement A Tracking Method**
  - When you begin to perform analysis, this rule will make more sense than it probably does now but it is important to keep in mind.
  - When you manipulate vector spatial data files, you will frequently be generating a copies of the data that have "added value".
  - When you manipulate raster data files, you will frequently be performing a process multiple times before you get the outcome you wish.
  - Therefore, it is important that you have a naming convention that will allow you to keep track of these sequential files.
  - Example: you may start with a shapefile of town_boundaries which only has one attribute (e.g. the name of the town). Performing a process called a Join results in new attributes (population of each town, zip code, the number of grocery stores in each town, etc) being appended to the table. Once the join is completed, you would export the entire file to a new file (e.g. town_boundaries_2) in order to permanently retain the new attributes.
  - Because you have implemented a tracking method it will be easy to distinguish the latest file you are working with.