Lesson 1: Reviewing the Basics of Geospatial Data

## INTRODUCTION

This lesson is designed to help you learn about important aspects relating to the geospatial data organization. You will gain an understanding of the various data formats and how this data can be organized and used to create different types of databases for effective use of geospatial data.

## LESSON OBJECTIVES

By the end of this lesson, you will be able to:

1. Define basic data acquisition and management terms.
2. Describe the basics of geospatial data organization.
3. Format raster and vector data.
4. Differentiate between a hierarchal, network, and relational database.

## LEARNING SEQUENCE

|  |  |
| --- | --- |
|  | |
| Required Reading | Read the following:  Reviewing the Basics of Geospatial Data   * Data Formats * Data Organization * Databases |
| Assignments | Complete the following assignments:   * Lab 1: Data Acquisitions * Module 1 Quiz: Reviewing the Basics of Geospatial Data |

## INSTRUCTION

## Data Formats

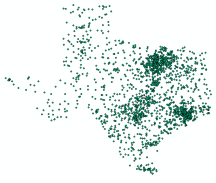
### What are Data Formats?

There are three commonly used geospatial data formats. They are vector, raster, and attribute. As this topic is only a review, these three geospatial data formats will be briefly discussed to refresh your memory.

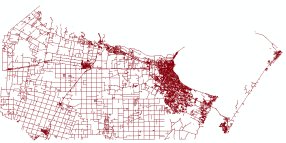
### Vectors

A vector geospatial data format represents discrete objects such as a police station. A vector is composed of coordinate pairs and associated attributes. Vectors can be stored in many different file formats, for example a shapefile or TIN. The three basic types of a vector are point, line, and polygon.

#### Point

A point vector object is composed of a single set of coordinates, for example, latitude and a longitude.

#### Line

A line vector object is composed of at least a second ordered set of coordinates where the first coordinate specifies the start position of the line, and the second coordinate represents the end position of the line. The start and end positions are known as nodes. If the line changes directions anywhere between the start and end nodes, then another coordinate set is included and is known as a vertex. The vector data format does not store the line drawn between the coordinate sets; instead, the computer draws the line when the data is to be rendered on the screen.

#### Polygon

A polygon vector object is composed of ordered sets of coordinates where the first and end coordinate are one of the same so that the polygon closes. The first and end point are known as nodes, and any set of coordinates between are known as vertices. A polygon has both an interior portion and an exterior portion, where the interior portion is the area encapsulated by nodes and vertices. Similar to the line format, the vector geospatial data format does not store the boundary line and filled area of the polygon, instead, the computer renders those lines and areas.

## Triangulated Irregular Network (TIN)

### What is a TIN?

A triangulated irregular network is vector-based data constructed by drawing triangles between points. The result is a surface that is rendered as a connected network of triangles. The data structure only stores the vertices of the triangles, and the computer renders the edges, or lines, between the points, and the faces which make up the area inside the triangles. The triangulated irregular network typically represents elevation values or topographic data, but of course can represent any attribute or phenomenon. The image provided is an example of a triangulated irregular network showing topographic information.

[](//upload.wikimedia.org/wikipedia/commons/d/d9/2003-3d-hawaiian-islands-usgs-i2809.jpg)

### Rasters

A raster represents continuous objects such as temperature and elevation. A raster is composed of a set of cells in a grid pattern. Each cell holds a single value. Examples of rasters are images, Geo-TIFFs, digital elevation models, and Landsat imagery.

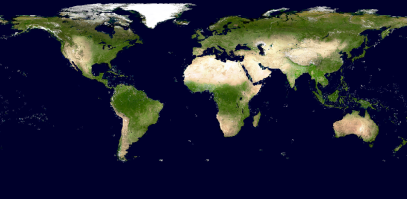
### Images

Images can refer to many different type of imagery such as satellite imagery, aerial imagery, or imagery from any other sensor such as a thermal image. When these images are brought into a GIS, they can be used for many useful purposes such as support for maps as a background image, analysis, land-use classification, or 3-D renderings.

[](//upload.wikimedia.org/wikipedia/commons/b/b1/Flickr_-_Official_U.S._Navy_Imagery_-_Aerial_of_Wakuya,_Japan._(1).jpg)This image shows a destroyed portion of Japan from a tsunami. These images can be used on map layouts to provide context to the map reader.

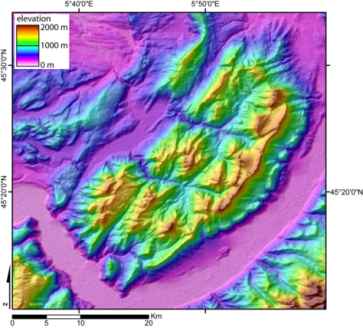
### Geo-TIFF images

A Geo-TIFF is an image allowing geo-referencing information to be embedded into a TIFF file. It can provide information such as projections, coordinate systems, and spatial reference information that may be useful to other GIS users. With this geo-referencing information, a user can digitize from the photo or bring it into the GIS and have it automatically appear where belongs in the world.

This Geo-TIFF image shows a majority of the world. When brought into a GIS, this TIFF is automatically placed in the correct location, and if you digitize from the TIFF, then the digitized objects will be placed in an earth-based coordinate system.

### Digital Elevation Model (DEM)

A digital elevation model, more commonly referred to as a DEM, is a 3-D representation of the terrain. It is a grid of regularly spaced elevation samples. A DEM can also be referred to as a digital surface model if it shows the bare earth elevation or a digital terrain model if it shows the bare earth elevation including the elevation of features on the earth, such as buildings and trees.

[](//upload.wikimedia.org/wikipedia/commons/5/5a/DEMChartreuse.png)This image is an example of a digital elevation model where different elevation values are rendered in gently varying colors. When combined with shadow modeling, digital elevation models can provide an attractive and easy to understand rendering of the terrain.

### Landsat

Landsat is a series of satellites that are constantly monitoring earth and its resources. It takes photographs of the surface and provides us with important satellite imagery that can be utilized in GIS.

[](//upload.wikimedia.org/wikipedia/commons/c/ce/Landsat-7.jpg)This image is a rendering of a Landsat satellite as it orbits the earth.

[](//upload.wikimedia.org/wikipedia/commons/c/c0/Landsat_bombinhas.jpg)This image was captured by a Landsat satellite. Notice it is a low resolution image, however, it allows us to easily identify physical and man-made features at a small scale.

### Raster vs. Vector

When choosing which geospatial data format to store your geospatial data in there are a few things to keep in mind with respect to advantages and disadvantages of each format. The raster format is a simple data model that is easy to store and easy to render. As a raster is a regular set of grid cells, it is easy to overlay multiple rasters if they have the same origin and grid cell size, which allows GIS analysts to perform some powerful overlay analysis.

It almost goes without saying that the raster format is best for storing digital imagery, whether it is a satellite, aerial, or terrestrial photography. The vector geospatial data format is very compact in regards to file size. A vector is great for network and linear features as it natively supports lines and can support connectivity. Humans are more familiar with vector data as these are the same graphic primitives that we draw with using pen and paper. Lastly, vector files can contain topology which allows simple or complex spatial relationships to be preserved even if a feature is moved.

## Attribute

### What are Attribute File Types?

An attribute is a record of a non-spatial characteristic. Attributes may or may not be related to geospatial features. Attributes are often rendered as tables and may be stored alone in a simple text file, or included in a complex database. There are four types of attributes: ordinal, nominal, interval, and ratio.

### Ordinal Attribute

An ordinal attribute ranks values. These ranked values can be descriptive such as high, medium, or low, or numeric, such as a range from 1 to 5. An ordinal attribute does not have any form of scale which means that we cannot determine the magnitude of the difference between one ordinal attribute value and another. If we consider high, medium, and low as three descriptive ordinal attributes, we cannot say that high is twice as high as medium and medium is twice as high as low. We can only infer that high is greater than medium and medium is greater than low.

|  |  |
| --- | --- |
| **Movie** | **Star Rating** |
| **A Knight’s Tale** | **5** |
| **Dark Shadows** | **2** |
| **Avengers** | **4** |
| **The Other Boleyn Girl** | **1** |

### Nominal Attribute

A nominal attribute provides descriptive information without specifying in order, size, or any other quantitative information. Examples of nominal attributes would be colors, names, or driver’s license numbers.

|  |  |
| --- | --- |
| **Name** | **Favorite Color** |
| Greg | Blue |
| Shaniqua | Yellow |
| Claire | Green |
| John | Blue |
| Sierra | Black |

### Interval Attribute

An interval attribute ranks both order and absolute difference (magnitude). It is often recorded as real numbers and has an arbitrary 0 point. An example of this would be if we looked at temperatures for three different cities, the temperature values are represented as real numbers. However, as the temperature is recorded in degrees Fahrenheit, these numbers are considered to be of interval type because 0°F is an arbitrary 0 point.

|  |  |
| --- | --- |
| **City** | **Temperature (°F)** |
| Corpus Christi | 98 |
| Dallas | 85 |
| Wheelock | 65 |

### Ratio Attribute

The ratio attribute ranks both order and absolute difference (magnitude) just like the interval attribute. Also, ratio attributes are often recorded as real numbers. The difference between interval and ratio is a ratio attribute has a natural 0. If we look at three students and their ages, and the age of a person has a natural 0 starting point, which is considered to be the day of their birth we can state that a 10-year old is twice as old as a 5-year old.

|  |  |
| --- | --- |
| **Student** | **Age** |
| Jim | 15 |
| Sally | 16 |
| Jessica | 14 |

### Example: Attributes

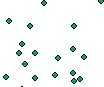
So let’s consider an example of how all of this comes together. In reality, we have a house at a particular location. The house is digitized into a spatial database and is represented as a point with a pair of geographic coordinates. That point is related to an attribute table which has four attributes named: house number, size, color, and cuteness.

First consider the cuteness attribute. The cuteness attribute records how cute the house is. There are three unique values of very, average, and not. While there is a definite ranking in these attributes, we cannot tell the magnitude between say very cute, and not cute. Because we do have a rank but not a magnitude, the cuteness factor is an ordinal attribute.

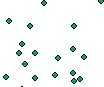
Moving to the color attribute, we have the three values of red, white, and yellow. As these attributes only describe the color of the house and do not provide any rank or magnitude information, this column is considered to be nominal.

Next, we have the size of the house with three different values of 5,000, 6,000, and 1,000. These values provide both rank and magnitude as we are definitely able to say that a 5,000 square foot house is five-times larger than a 1,000 square foot house. Therefore, the size attribute is considered to be interval. Now we come to the house number specifying where it sits on a street.

If we consider that a house number serves to differentiate its location from other houses on the street, they would consider it to be a nominal attribute. A case can be made for specifying a house number as an ordinal, interval or ratio, however, it does not make sense to consider the house number as being a way to rank the houses or specify how far apart they are from each other, therefore, nominal would be the best choice.



[](//upload.wikimedia.org/wikipedia/commons/3/34/Blades_House.JPG)





**Ordinal**

**Nominal**

**Interval**

**Nominal**

## Data Organization

### Why Organize Data?

Keeping data organized is a huge factor in project management. The speed at which a project is completed can be sped up significantly by merely keeping data organized. When it is easy to find data, the process is sped up for both you and other involved parties. Poorly organized data not only slows down projects but also frustrates everyone at the same time. Additionally, as a project is passed around, it makes it easier for the next person to pick up where the previous person left off if it is well organized.

### Ways to Organize Data

### There are many ways to organize your data for a project. At a minimum, you should create specific folders to sort your data, for example, placing all of your shapefiles for a project in a single-parent directory, or even going further in separating the shapefiles into sub folders based off of theme. You can also copy items to keep the original raw data without any edits in case you need to revert back to the original data set.

### Items that are no longer used such as intermediate output from tool runs should be deleted instead of kept. If you have a large amount of data and tools you may also wish to consider using a database to organize the information. Databases provide many advantages over more traditional format such as a shapefile and should be used if at all possible. And lastly, you should rename items to make recognition easier. The following methods can be used to organize data.

#### Files

Folders allow for categorization of data and improve your ability to navigate through and find the data you’re looking for. In addition to separating data by category, folders can also sort different days’ worth of work allowing the user to track the changes over time.

#### Copying

### Copying and creating multiple copies of files allows for tracking changes. This will also allow for multiple versions of the file or project if you are trying out different things. The nice thing about copying items is that if you make an error it makes it easier to go back to the version without having to start completely over.

### Copying items also provides you with a backup your project and it is recommended that you take these copies and move them to a different physical drive or to online storage to protect against failure of a single stored source. You can also use a version control system, such as Subversion, GitHub, or CVS.

#### Deleting

You should always be careful when deleting items; however, deleting items really helps to remove clutter for project. If you decide you do not need specific data than deleting it allows for more storage space for new files and keeps up any unwanted or unnecessary data. Additionally, deleting items helps the user from confusing and mixing projects or files.

#### Renaming

Renaming files is perhaps one of the easiest and most helpful things you can do when organizing your data. You should rename files to something the user is more likely to understand and remember. Operating systems today allow for filenames consisting of hundreds of characters in length. You should be verbose and descriptive when naming files. You will thank yourself later. Some downloaded files have hard to decipher names such as TX\_5000.SHP. While these filenames may have meaning to the creators, they are difficult to decipher, understand, and remember for the users. Feel free to rename these files to something more easily recognizable.

#### Databases

The creation of a database allows for the organization of information in a single structured location. Databases can do many powerful things such as keeping specific layers together, relating data sets to each other in complex ways, and enforcing formatting rules. While databases are little more complex than folders and files, they provide many benefits when organizing your data and should be used whenever possible.

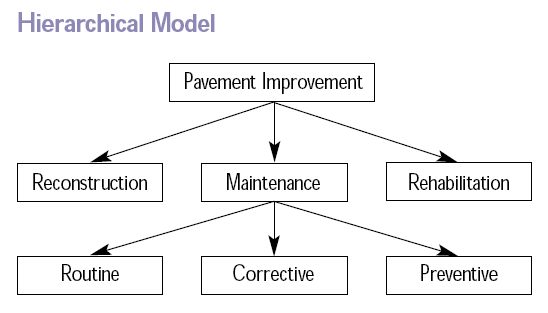
## Databases

### What is a Database?

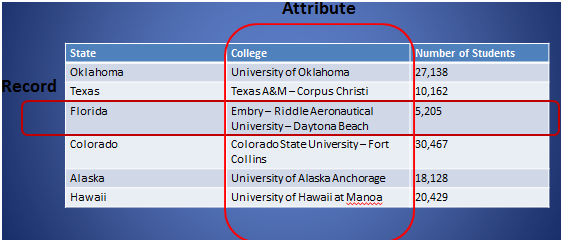
A database is a set of data that is structured and stored in a computer. Databases excel at organizing and manipulating data. There are three different types of databases commonly used today: hierarchical, network, and relational. Please note that there are other types of databases; however this lesson will only cover these three initially, then only focus on the relational database as it is the most widely used database technology.

### Hierarchical Database

## A hierarchical database structure organizes data and its attributes into a tree like structure. It uses common parent-child relationships. In a parent-child relationship, parents can have multiple children; however a child can only have one parent. In this structure, each record is represented as a row while each attribute is represented as a column.

[](//upload.wikimedia.org/wikipedia/commons/4/45/Hierarchical_Model.jpg)

Consider a database that holds information about pavement improvement. In this database, the pavement improvement table is the parent of three child tables. The reconstruction and rehabilitation tables do not have any children. The maintenance table has three children thereby increasing the depth of this hierarchical structure.

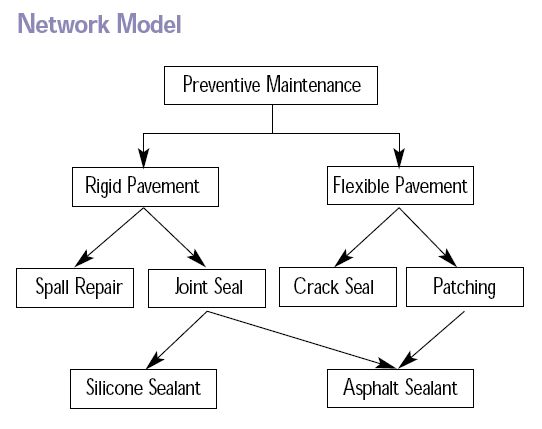
If we look at a single table in a hierarchical database, it is arranged in rows and columns. The columns are attributes, and the records are the rows.

## Network Database

### Network Structure

A network database uses a network structure to store the information. The network database is designed for client/server architecture. This network structure allows for data relationships to be created and also allows for a single attribute to point to multiple other attributes.

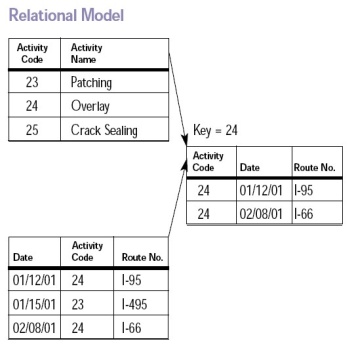
As an example, if we look at the preventive maintenance network structure, we can see that as we work down to the attributes. Each attribute points to other attributes.

[](//upload.wikimedia.org/wikipedia/commons/3/3c/Network_Model.jpg)

## Relational database

### What is a Relational Database?

A relational database uses a relational structure of information that collects data into table structures. These tables can be related to each other using intermediate tables which handle the relationships. Relational structures are the most common structure used today in databases and are often chosen over the other forms of databases. In the relational database structure, it consists of numerous tables with each table having its own unique identifying attribute used to relate tables together.

[](//upload.wikimedia.org/wikipedia/commons/8/85/Relational_Model_2.jpg)In this relational database there are three tables. Each of these tables has an activity code which allows each of the three tables to be related to each other based on the common attribute. Relating tables together using these common attributes allow for information to be logically separated into different tables. This serves to reduce the duplication of information and the database size among other advantages.

## SUMMARY

There are three commonly used geospatial data formats: vector, raster and attribute. You learned about triangulated irregular networks and attribute file types including ordinal, nominal, and interval/ratio that are used to record non-spatial characteristics of features. In this lesson you learned about the importance of creating files, making copies essential for tracking changes including the renaming of files as you prepare data that will be used to create databases. This process will keep complexity to a minimum and increase the efficiency of locating needed data within a well-defined database.

## ASSIGNMENTS

1. Lab 1: Data Acquisitions

2. Module 1 Quiz: Reviewing the Basics of Geospatial Data