Lesson 5: Creating Geospatial Data

## INTRODUCTION

## In this lesson you will learn about creating geospatial data. You will also learn how to perform coordinate transformation. The concept of digitizing is explored including automatic and manual digitizing including feature generalization and the methods used in this process.

## LESSON OBJECTIVES

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

1. Create a new Geodatabase to store digitized features.

## LEARNING SEQUENCE

|  |  |
| --- | --- |
|  | |
| Required Reading | Read the following:  Creating Geospatial Data   * Geospatial Data * Coordinate Transformation * Digitizing * Manual Digitizing * Automatic Digitizing * Feature Generalization |
| Resources | View the following:   * Manual Digitizing |
| Assignments | Complete the following assignments:   * Lab: Creating Geospatial Data * Quiz: Creating Geospatial Data |

## INSTRUCTION

**Geospatial Data**

## Creating Geospatial Data

There are many reasons why you might want to create your own geospatial data. The main reason is that the data you need has not been created by anyone else. Even though GIS and digital data have been around for a long time, there is still a large need for the creation of new data covering many different facets of our world at many different scales. A couple of additional reasons that you may want to create geospatial data are that the data you need is on a paper map and needs to be converted to digital format so it can be used in your GIS software, and the data available may be a different scale than you need. For example, let’s say you want to show a map of where all the state parks are in Virginia. However, all of the park data sets are extremely detailed, and look very poor at a statewide scale because of the intricate detail of the park boundaries. In this case, you would need to create a new park dataset suitable for display at a smaller scale, or, simplify the existing dataset so that it can display better at the smaller scale. Another reason to create new geospatial data is the possibility that the data available is too old for your purposes and you need an updated version of the data. There are many other reasons why you would want to create geospatial data, but these are the most common reasons that you should be aware of.

|  |
| --- |
| **Why Create your own Geospatial Data?**   * Data you need has not been created * Data you need is on a paper map and needs to be converted to a digital format * Data available is at a different scale than you need * Data available is too old for your purposes |

## Two Types of Data Sources

There are two types of data sources that you can derive digital geospatial data from: hardcopy and digital.

Each of these data sources has their own positive and negative aspects as a way to store data. Let’s look at these aspects, starting with hardcopy data.

## Hardcopy Data

Hardcopy data, such as data stored on paper as maps, was the most common medium of storage for GIS until the 1980’s. The positive aspects of hardcopy storage is that it is a stable medium which means that it cannot get a virus, run out of battery power, or crash, like a computer. Hardcopy data is also cheap to produce, is easily portable, and does not require any specialized hardware or software to use. This provides a very low barrier for use, and makes a paper map easy to use and understand by a very wide range of users. Hardcopy data is semi-permanent so long it must be properly stored. The negative aspects of hardcopy data are that it is not reusable because it cannot be easily updated. A hardcopy map, for instance, is a static representation, and can only be updated by printing over the existing map, which is often not desirable.

|  |
| --- |
| **Key Facts**  Hardcopy data is…   * stable * cheap to produce * does not require hardware/software to use * semi-Permanent * portable * familiar to many users/audiences * reusable * not easily updateable |

|  |  |
| --- | --- |
| Hardcopy data, such as data stored on paper as maps, was the most common medium of storage for GIS until the 1980’s. The positive aspects of hardcopy storage is that it is a stable medium which means that it cannot get a virus, run out of battery power, or crash, like a computer. Hardcopy data is also cheap to produce, is easily portable, and does not require any specialized hardware or software to use. This provides a very low barrier for use, and makes a paper map easy to use and understand by a very wide range of users. Hardcopy data is semi-permanent so long it must be properly stored. The negative aspects of hardcopy data are that it is not reusable because it cannot be easily updated. A hardcopy map, for instance, is a static representation, and can only be updated by printing over the existing map, which is often not desirable. | **Key Facts**  Hardcopy data is…   * stable * cheap to produce * does not require hardware/software to use * semi-Permanent * portable * familiar to many users/audiences * reusable * not easily updateable |

Digital Data

Digital data has become a common storage medium for GIS data since the 1980s as computers became more widely used and storage capacities grew larger and prices dropped significantly.

|  |
| --- |
| **Key Facts**  Digital data is…   * + stable   + cheap to produce   + semi-permanent   + flexible   + easily updateable   + not portable   + requires hardware and software to use   + may be tied to specific systems that may deprecate |

### Advantages of Digital Data

The positive aspects of digital data are that it is stable, so long as it is properly maintained, backed up, and occasionally transferred to a modern storage medium. Digital storage is now extremely cheap and can store massive amounts of information in a very small space. Digital data is semi-permanent. What is meant by that, is that the data structures must be set up to allow for archiving of historical values information, however, this is often not the case, therefore previous data is often overridden when values are updated and there is no way to see the previous version or value. Digital data is extremely flexible. Digital data can easily be updated and can be used for many different purposes.

### Disadvantages of Digital Data

Negative aspects of digital data are that it is not portable without specialized hardware and software. This significantly raises the barrier of entry for users, and the hardware and software may not be easily understood by the users either. Additionally, the data may be tied to specific systems that may deprecate over time, so there is a need to continuously update the data into newer hardware and software systems.

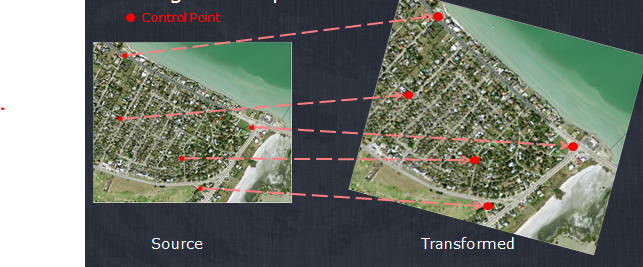
**Coordinate Transformation**

## What is Coordinate Transformation?

Coordinate Transformation is the process used to bring spatial data into an earth-based map coordinate system using a series of control points.

For instance, imagine you are in an airplane and took a bird’s eye view of Earth. You can copy that picture from the camera to the computer and attempt to place it in your GIS software. What would happen is that the image will show up not in the right place on Earth, but will, instead, most likely be placed at the intersection of the equator and the Prime Meridian. This is because the aerial photo you took is not referenced to any location on Earth, in fact, the picture has its own arbitrary coordinate system which starts from 0, 0 point, typically in the lower left-hand corner, and increases in a positive direction along the X and Y axis.

That is why, when you add a picture to a map, it places it at the intersection of the equator and Prime Meridian, because that is the 0, 0 point in the latitude and longitude system. In order to make the aerial photo show up in the right position on Earth, we must perform a coordinate transformation. The idea is that you identify locations that are in your picture on the ground, and get the coordinates for those locations. You then match up the known coordinates of the positions on the ground with where it is located on your picture. Once you have chosen enough control points, the computer software can perform a transformation, by placing the picture where it belongs on Earth. During the transformation, the picture is often stretched, rotated, and deformed, to match the location of the control points on the earth and on the photo.



Performing Coordinate Transformation

In order to perform the coordinate transformation, you must use considerable care when choosing which control points to use. The control points are used to transform the data from the source coordinate system, to the map coordinate system. The source or digitizer coordinate system in the case of our example is the coordinate system used in the digital camera. The map coordinate system is the coordinate system that you wish to transfer to, such as the state plane coordinate system, or latitude and longitude. Therefore, in order to perform the coordinate transformation, we needed to identify control points. One set of control points from the source, or digitizer and one set of control points from the map.

## Control Point Criteria

When choosing control points on both the source and map, you should choose control points that have the highest feasible coordinate accuracy. This means, that if you’re looking for an accuracy of 1 foot, you control point should be no larger than 1 square foot in size. Additionally, for the control point, the accuracy of the control point’s location should be at least as good as the desired overall positional accuracy. That means that if you want centimeter level accuracy, then you should not measure the location of your control point to the nearest 5 meters. And last, your control points should be evenly distributed throughout the area. By distributing the control points evenly throughout the area, you are providing sufficient coverage to provide a good transformation across the entire area. This in contrast to only choosing control points in one corner of the data set, which in that case, only the corner of the data set will be accurate, and the rest of the data set will contain significant amounts of error.

**Control Point Criteria**

1. Should provide highest feasible coordinate accuracy
2. Accuracy should be at least as good as desired overall positional accuracy
3. Should be evenly distributed through the data area

## Sources of Control Points

There are two common sources of control points. The first source is a land survey performed by a licensed land surveyor. Land surveyors are experts at measurement, and can provide very accurate and precise control points based on existing or new benchmarks that are placed around the nation. The second data source is an existing map that covers the same location as the source data set. If the existing map is already registered, then you can pick the coordinates from the map as your control points. When choosing control points off of an existing map, good choices are permanent features that are easy to see from the air, such as concrete corners, bases of buildings, or aerial photo targets which are those white crosses you often see painted on the highway. Poor choices will be items that move, or are difficult to see from the air such as cars, trees, or street paint.

* Land Survey – base control points on benchmarks or set new control points
* Existing Maps – base control points on existing maps

Land Survey

Bases control points on benchmarks or set new control points

Existing Maps

bases control points on existing maps

|  |  |
| --- | --- |
| Good Choices | Poor Choices |
| * + - Concrete corners     - Base of buildings     - Aerial photo targets | * + - Cars     - Trees     - Street Paint |

**Digitizing**

## What is Digitizing?

Digitizing is the process of converting coordinates of features from a data source, such as a paper map, into a digital format. Digitizing can be thought of as tracing an object. The general idea is that we will use an input data source, such as an aerial photograph rectified to earth-based coordinates, and then use an input device to trace and record the features from the map as a vector feature. There are two methods of digitizing: manual digitizing and automatic digitizing. Manual digitizing is when a person digitizes the features from the source data manually. This process is often time-consuming, tedious, and prone to human error, however, it allows the human to intelligently determine where to digitize, and what to digitize. The second method of digitizing, automatic digitizing is where a computer program is trained to identify features on the input data source, and then automatically traces those features. Automatic digitizing can save a significant amount time; however, the features need to be easily identifiable on the input data source. Training the software to identify the features may take quite a bit of time, and a human may still need to clean up errors in the automatic digitizing results.

## Anatomy of a Digitized Object

When an object is digitized into the computer from the source data set, it is composed of two parts: a node, and a vertex.

A node is a start or end point of a line segment, or a point.

A vertex, is an intermediate point of a segment and must exist between two nodes.

|  |  |
| --- | --- |
| As an example, if we look at the digitized object on the slide, we can see that there are four nodes, and five vertices. As we know that nodes determine the end and beginning of a segment, and vertices are continuations of the same line segment, we can identify four separate digitized line segments in this figure. Starting at the bottom, we have a node, followed by four vertices, that terminates at the base of the triangle. That is one line segment. The second line segment is between the two nodes on the small leg of the triangle. The third line segment, are the two nodes on the short leg of the triangle, that go through the vertex at the other end of the triangle. The fourth and final segment is the short line that offshoots from the top node of the triangle to the end of the short line. |  |

## Common Digitizing Errors

When digitizing manually, there are two common digitizing errors. The first common digitizing error is known as an overshoot. An overshoot is where lines cross over existing lines and nodes where they should’ve actually connected. For instance, imagine if we are digitizing pipelines off of a map and we have two pipes that connect in a T intersection. However, when digitizing, the pipe that is forming the stem of the T does not connect exactly with the other pipeline; instead, it passes over and through it in the digital data making a cross instead. This is an example of an overshoot. An undershoot is when nodes do not quite reach the intended line or node. Going back to the pipeline example, if the two pipelines do not intersect, because the second pipeline was terminated just short of the pipeline was supposed connect to, this would be considered an undershoot.

|  |  |
| --- | --- |
| *Overshoot* lines that cross over existing lines or nodes where they should have connected  *Undershoot*  nodes that do not quite reach the intended line or another node |  |

## Digitizing Process

When performing digitizing, you should follow these five steps.

Step 1: Create or use an existing data set to store the digitized data. You

should select the geometry type of the data set.

Step 2: Load the source data into the GIS program.

Step 3: Register the source to Earth-based coordinates if it is not already.

Step 4: Manually, or automatically, digitize the features into the digital data set.

Step 5: Save the newly digitized features.

Manual Digitizing

Manual digitizing is the human guided capture of features from a map image or source. There are two methods of manual digitizing: on-screen digitizing, and hardcopy digitizing.

On-screen digitizing is probably the most common form of digitizing today and is where a source is scanned, or downloaded into the computer’s memory and loaded into digitizing software. The mouse and keyboard are used to digitize features while the sources feed on the screen.

Hardcopy digitizing is where the original source is taped to a digitizing table which is connected to a computer. The digitizing puck, which is similar to a mouse, is used to digitize features from the hardcopy source into the computer. The purpose of the digitizing puck is to feed coordinates into the computer, based on its location on the digitizing table, which would be the source’s coordinate system.

Watch, [Manual Digitizing](http://www.youtube.com/v/t9NnO_T2fuw?version=3&hl=en_US&rel=0) (1:46), a demonstration of on-screen manual digitization.

Video Transcript: Manual Digitizing

This video is an example of on-screen manual digitizing. Note that as we are digitizing the centerlines of the roads, it is being stored into a line vector data set. Additionally, after the road is digitized, it provides an attribute with the name of the road. As we digitize the houses, it is being stored into a separate polygon vector data set. It is also being attributed after the polygon is completely digitized. At the end of the digitizing process, do not forget to save your work, otherwise all of your digitizing will be lost.

## Manual Digitizing: Positives and Negatives

**Positive Aspects of Manual Digitizing**

Manual digitizing is typically sufficiently accurate as humans can interpret maps easier than computers. As digitizing is not typically a job that requires highly skilled operators, the training period is often short and the labor is cheap.

**Negative Aspects of Manual Digitizing**

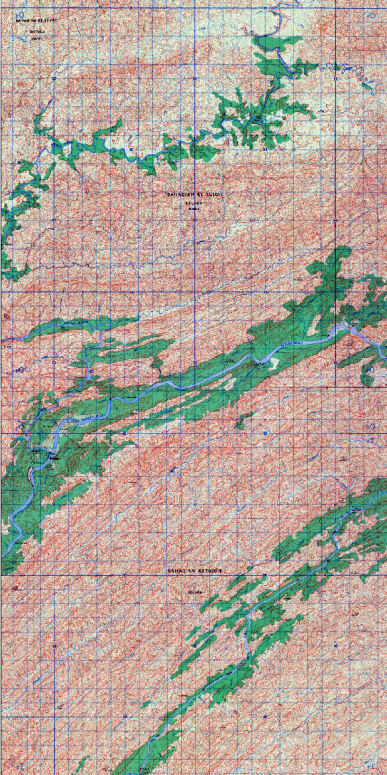
There are a couple of negative aspects to using manual digitizing. For instance, the map scale impacts the accuracy. Different operators may work better or worse at different scales, as some maps may have such fine detail that the operator has a hard time accurately digitizing the intended feature. The hardware that the operators are using may also impact accuracy, if the operators are not provided with capable hardware and software. Finally, different operators will digitize at different qualities and may interpret what is being digitized differently. This may lead to inconsistent digitizing results.

**Automatic Digitizing**

What is Automatic Digitizing?

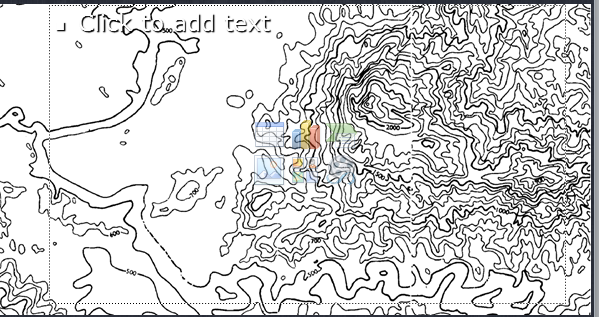
Automatic digitizing is the computer guided capture from a map image or source. This is different from manual digitizing, where it is a human guided capture. The way in which the computer performs automatic digitizing is that it uses algorithms to identify and digitize the features. While these algorithms are often quite powerful, it often requires an initial training period that may take time. Some of the algorithms available in automatic digitizing software can learn how to identify the features and are supposed to trace based on full color scans, instead of requiring an initial binary classification.

## Example: Automatic Digitizing

 Here is an actual example of automatically- digitizing contour lines from a contour map. What you are seeing on the slide, is actually three contour maps that have been mosaicked, or pasted, together. As you can see, there are quite a few red contour lines on this map, which would take quite a long time for someone to manually digitize. In an effort to save time, we will automatically digitize the contour lines from this map. This map is a good candidate for automatic digitizing because the contour lines are drawn in a contrasting red color, and no other features on the map are the color red. This helps the computer easily identify, and isolate the contour lines of the map.

### Step 1

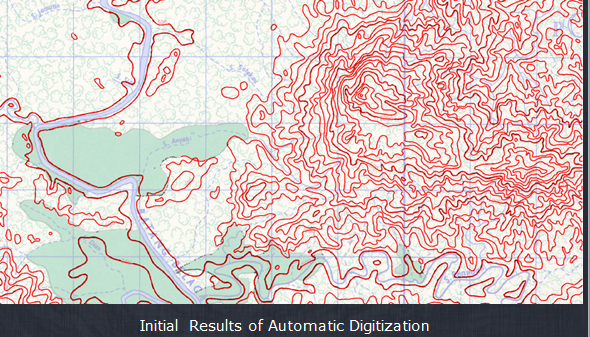
In order to assist the computer in isolating the contour lines, the digitized contour map has been classified into a binary image. Wherever there was red on the map, is now black, and where the color red did not exist on the original map, it is now white. This initial step in the automatic digitizing process and is critical for ensuring that the computer can easily distinguish features from the map. It is important to note, that other digitizing software does not need the user to isolate the features for it, as it can work with selected colors.

**Step 1: Isolate Contour Lines** 

### Step 2

The second step of automatic digitizing is to run the automatic digitizing algorithm. The red lines now shown on the map are the actual digitized vector lines. We cannot see the original raster representation of the contour lines, as they are underneath the digitized lines; this means that the automatic digitizing algorithm did a good job of tracing the contour lines.

Step 2: Run Automatic Digitizer



Step 3

The third step in automatic digitizing is to clean the results of step two. On this particular map, if the contour lines make sharp turns, the contour lines become disconnected. For instance, the blue contour line should connect with the red contour line that it touches toward the center of the screen. However, as this is a significant change of angle, the automatic digitizer considers this to be separate contour lines. Therefore, it is up to the user to either: modify digitizing parameters so that this may be caught, and run step two again, or, manually connect the contour lines.

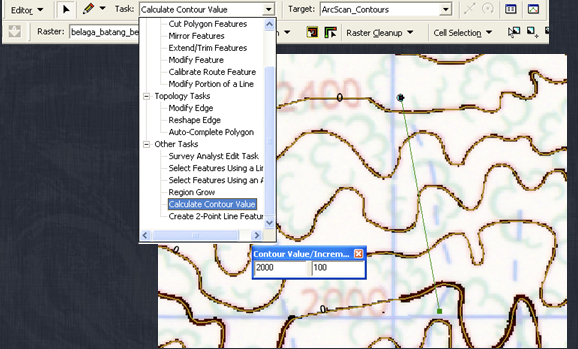
Cleanup Task Connecting Disconnected Contours

|  |  |
| --- | --- |
| Example of Cleanup Task Connecting Disconnected Contours |  |
| Example Cleanup Task Connecting Disconnected Contours  This is what the corrected contour line looks like. In this case, the user opted to manually connect the contour lines. Now that that error has been corrected, the user would continue to review the automatic digitizing results to find the other errors, and correct those. |  |

### Step 4

The fourth step is to assign the contour line values into the attribute table of the newly digitized contour lines. In this case, a tool was downloaded that makes it easy to assign multiple contour lines at once, by drawing a line across multiple contour lines, and then specifying the first contour line it crosses contour value, and how much it should increment each additional contour line that passes over. So, in this case, the contour line at the bottom would be assigned a value of 2000, the line above it 2100, line above that in the line above that 2220 and 2300 respectively, and finally, 2400 as the last, top contour line is crossed.

**Step 4: Assign Contour Line Values**



This is the result of the automatically digitized contour lines with the contour values labeled.

**Completed Contour Lines** SLIDE 10

Finally, this shows the completed digitized contour lines overlaid on top of the original contour map. As you can see, the digitized contour lines follow the original contour lines extremely close and it does a good job of saving the manual digitizer time.

|  |  |
| --- | --- |
|  |  |

## Automatic Digitizing: Positives and Negatives

**Positive Aspects of Automatic Digitizing**

Automatic digitizing is good for large projects that would take a large amount of staff and time to complete. There is no initial operator error but the results will require manual editing, as we just saw. Automatic digitizing can work well with a large number of elements on a map, so long as those elements are easily distinguishable by the computer.

**Negative Aspects of Automatic Digitizing**

There are a few negative aspects of using automatic digitizing. One negative aspect of automatic digitizing is that the software can be very expensive; however, if the project is sufficiently large, the expensive software may ultimately be cheaper than the labor that would be required for manual digitizing. Automatic digitizing can also only translate and trace, but cannot interpret what it is actually looking at. Automatic digitizing may be more susceptible to miss digitizing areas in the map where the scanner messed up or had a smudge.

For most of your digitizing work, if it is sufficiently small in scope, manual digitizing will be the preferred method. However, if the digitizing work covers an extremely large area and the source has easily distinguishable features, automatic digitizing can provide a large time and money savings.

**Feature Generalization**

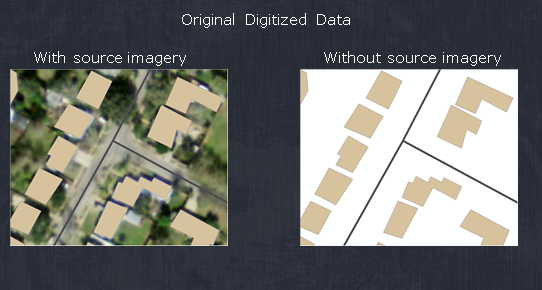
## What is Feature Generalization?

We should first step back and remind ourselves that maps are simply abstractions of reality. That is, they are someone’s viewpoint of how the world is organized. While we would like to store infinite detail about reality in geospatial data sets, it is simply not possible. Constraints on an infinite representation are the size of the map medium, insufficient resolution of sensors, finite disk space, scale at which we are representing reality, and so on. Additionally, you may not need to represent all aspects of reality for your mapping purposes.

Our inability to store infinite detail of reality and geospatial data sets means that we must generalize the features that we are storing. Feature generalization is the approximation and simplification of real features represented on a map. The main purpose of feature generalization is to aid in map legibility by reducing the complexity of the features being shown. What is most important to remember about all feature generalization, is that a properly generalized feature, should still maintain the same character of the original objects. That means, for example, if we reduce the number of curves in a stream, we should still be able to identify that feature as a simplified version of that stream.

Four Methods of Feature Generalization

In order to discuss the four methods of feature generalization, let us first see what the source data was. The source data for this digitizing project was a digital orthophoto quadrangle quad. Houses could be easily identified on these photos, and the user wanted to digitize the footprints of the houses and roads. The original source data is shown on the left with the digitized house footprints laid on top. On the right, we see the digitized data without the source imagery behind it. We will now focus on only the digitized objects and the ways in which we can generalize those objects to aid in map legibility when we display this information on maps.



## Feature Generalization Methods

View the following examples of feature generalization methods: fused, simplification, displaced, omitted and exaggeration. It is important to reference, the original digitized data image in order to gain an understanding of how each method affects the image.

### Fused

The process of fusing is to combine the geometry of adjacent features that are close together. In this example, we have fused together eight houses so that the fused data now only has nine building structures, instead of the original 13.



### Simplification

To simplify features, you remove vertices where the feature shows a reasonable amount of complexity. For instance, looking at the house south of the intersection, we see a sort of stair-step pattern in its construction. If we remove the stair-step’s, and square off that corner, we still maintain the character of that house, as the relative location and shape are still intact.

### Displaced

Here, the houses are moved closer to the streets. The houses could also have been moved to be further away from the streets. In either case, what is important, is that the relative positioning of all the features involved still be maintained.



### Omitted

Here, features are simply removed from the data set. You may want to do this for lakes, for instance, where a large lake is surrounded by multiple small, insignificant lakes. If we remove the small insignificant lakes, keep the large lake, and zoom out, the map reader can still understand that there is a water body in that location.



### Exaggeration

Exaggeration is, perhaps, the most common method of feature generalization. For instance, on a roadmap for state, the roads are greatly exaggerated so that they can be seen on the map. If the roads were drawn to scale, they would probably be so thin, that the printer cannot print the map, and/or the reader could not read the thin line. Therefore, by exaggerating the road, the original characteristic of the road is still maintained, and a map reader can easily see the road features of importance on the map.



## Broad Directives for Generalization

Now that you are familiar with the methods for feature generalization, it is important to review some broad directives.

1. Observe the map purpose at all times.

Do not include pointless details or features and vice versa. Additionally, the degree of simplification must be related to the purpose. For instance, if you are creating a map of a river for a kayaking club, you would not want to overly simplify the river, as all of the intricate bends and features of the river are what the kayakers really want to see.

1. Scale has a bearing on the need for simplification.

The scale that the map is being produced at should have a bearing on the need for simplification. If, at the current map scale, you can easily see all the features on the map, then there is no need for feature generalization. Again, you should maintain the essential character of the feature and area. Even if the features have been fused, simplified, displaced, omitted, or exaggerated, the map reader should still identify those features.

1. Be consistent in treatment.

If you omit a large number of one type of feature layers on a map, but do not omit any of another type of feature layer on the map, the user may assume that the map layer with all of the omitted features is just as accurate as the detailed layer.

## SUMMARY

In this lesson you learned about how to create geospatial data. Coordinate transformation was explained and illustrated to teach you how to perform a transformation. You learned about manual and automatic digitizing and the benefits of each process. The four methods of feature generalization were explained with examples to help your understanding of the methods.

## ASSIGNMENTS

1. Lab: Creating Geospatial Data
2. Quiz: Creating Geospatial Data