CHAPTER TWO – DATA INPUT

As mentioned in Chapter 1, GIS must have capabilities for inputting data; various housekeeping functions to edit, store, and reshape data; analytical operations that manipulate the data into usable information, and output functions. In this chapter, we focus on data input.

There are three distinct phases to the data input process (Figure 2.1): In the first phase, the design phase, you identify and "conceptually" code all the needed features and attributes for your project. Next is the data acquisition phase, which involves acquiring the needed data from various agencies, storehouses, and organizations, and getting it into a format that your GIS program reads. Finally, in the data capture phase, you digitize hard-copy maps and data directly into your GIS and transform existing digital data into a format your GIS reads.

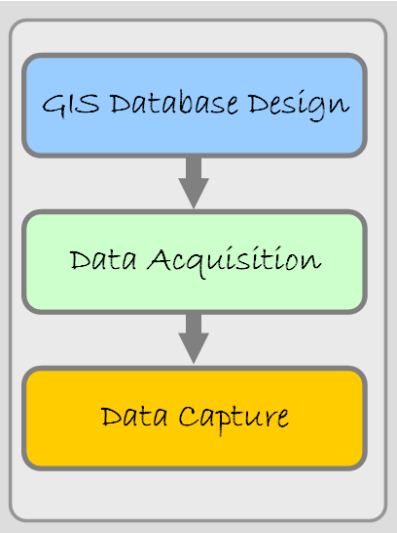

Figure 2.1: The three phases of the data input process.

PHASE 1: GIS DATABASE DESIGN

What is your goal or your research question? How should you proceed? You need to define your objective

at the very beginning. Having a well-defined research question, goal, or even multiple goals is the key to a successful GIS project because it guides the project's input, analysis, and output stages. Spend time and thought on the design of your GIS because good planning results in successful projects.

Start by thinking about the people, land, and the issues in your study. This has a direct bearing on what datasets (features and attributes) are needed. Next, think about how you will analyze the data. This could affect your choice of GIS software and your data model (vector or raster). Even envision how you would like to present your results. In other words, think through the entire project. If you are working for someone else, consider your employer’s purpose and understand how the company, agency, or organization functions.

It seems like everyone wants to use GIS for their projects today. It is both trendy and powerful, but you should ask yourself whether GIS is the appropriate tool. It may not be. To determine if GIS can help achieve your objectives, think again about the relevant features (land, people, and issues) of your study and whether these can be displayed geographically. Almost all features have geographic locations, but you must further decide if the feature's "geography" is important. One might be interested in population and economic statistics like income, population density, age, and ethnicity, but are you interested in how these variables differ across neighborhoods, across cities and towns, or across space in general? If location does not play a role in your study, then the data are "aspatial," and you should close this e-text.

If you can conceptualize your variables as features on a map or depicted within geographic boundaries like census tracts, ZIP Codes, or agricultural fields, GIS can aid your project. As an employee or a researcher, you may already have your goal or research topic chosen for you by your superiors. Your goal or research question might be something like "I want to know the best places to plant sorghum in Mauritania when the new dam goes in." Alternatively, if you are in a large planning department, you might have multiple goals, and then you must try to satisfy all the tasks, which might include updating and producing high quality maps, generating lists of residents that need to be contacted when a zoning amendment or a liquor license is requested, or determining which roads need resurfacing. It all starts with your goals. If your goals are fuzzy in your own mind, stop and sharpen them. Nothing in the design phase is more important.

Although all components of a GIS project need to be planned, including what software and hardware you will use and what procedures and people will guide your operation, this chapter focuses on data.

In this first phase of the input process, you determine what features and attributes are needed and how they should be coded. This starts with identifying each type of feature and their related attributes, but it goes beyond identification to include several important planning decisions. Figure 2.2 features a list of planning tasks. A discussion of each item follows.

Figure 2.2: Key questions to ask yourself in planning a GIS database.

1. Determine Your Features

What features are necessary? Think back to your project's goals. If you want to analyze a particular plant species' distribution, it may be necessary to have a feature devoted to the specific plant type. Equally important, however, are the other features—nearby plant species, soil types, climate conditions, land tenure practices, and landform conditions like slope and aspect. These other features, along with many others, play a role in the distribution of your plant. If you are developing a GIS database for a city's

planning department, you will want layers for many features including streets, parcels, parks, water, sewer, electricity, and buildings. In this case, the features are obvious.

The question of which features to use, however, may seem simple, but frequently it can be complicated. For example, you might want to interview hundreds of people in Modesto, California about their family income and quality of life. Is it appropriate to construct a layer with a point location at the home of each respondent, or would it be better to aggregate the responses of the individuals into neighborhood or census tract boundaries? In the first scenario, your feature layer might be called "respondents", and each point feature would be located at the home of an individual respondent. Attributes would be stored within each individual feature. In the second scenario, you would have a feature called "census tracts" and the aggregated responses of all the individual respondents would be contained within the appropriate census tract's data table. In other words, in the second scenario you would not have a respondent layer; you would have a census tract layer with aggregated responses. While the appropriateness question is hypothetical, you can see the ethical issue. You may not want to produce a map or provide the data to others that has the exact locations of your respondents if it includes sensitive or personal data.

Yet as a general rule, it is better to input data in its most precise and detailed form. In the income and quality of life example, you could create the more precise respondent layer and later aggregate (see Chapter 5) the individual respondents to census tracts for output purposes. The public will not see the detailed data. Benefits of having precise data come in verifying your census tract results, and in changing the resolution of your study. For example, what if—down the line—during the analysis portion of your project, you discover that the census tract boundaries are too coarse. It would be difficult, and perhaps impossible, to change or disaggregate the aggregated data into smaller, more detailed boundaries. If you have the respondent layer, however, one can quickly re-aggregate the data to a different and finer boundary feature like census blocks or a neighborhood unit.

2. Determine the Project's Spatial Extent, Scale, and Temporal Extent

You must determine the area and the period in which your project focuses. Sometimes it is obvious. If you

are working for a county planning department, chances are your extent is the county boundary. Cities, however, need to go beyond their boundaries to sphere of influence borders that may lay beyond the city limits because these regions have a direct impact on the city and this region may someday become part of the city. Other boundaries, especially those in research projects, are more difficult to establish.

Along with the project's spatial extent, you should think about an appropriate scale. There is a relationship between scale and detail (see Figure 2.3). Small-scale maps depict large territories, but they usually are less precise and may require that some reference layers be left out. Large-scale maps show smaller areas but comparatively include more detail. Although GIS allows one to zoom in at increasingly larger scales, data captured at a “small scale” become inherently inaccurate when zoomed in on. The desired scale affects both the amount and precision of the data to be collected and the scale at which the geography can be cartographically represented. There is more on this topic later in this chapter under Precision Errors.

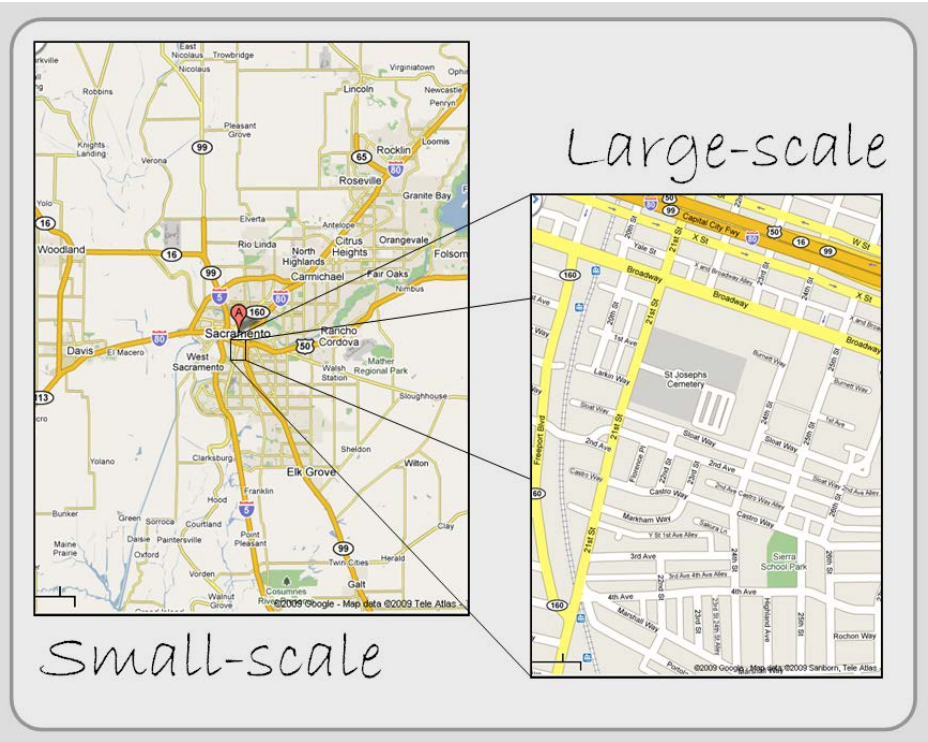

Figure 2.3: Map scale. Small-scale maps depict larger territories, but large-scale maps present more detail.

Similarly, you may want to define a temporal extent. Is time an important variable in your study? Most GIS

projects focus on the contemporary scene and ignore the past. With contemporary databases, the assumption is that the GIS database is up-to-date. That may be a major assumption since some of the layers might be quite dated. If, however, you want to determine how much an area has changed, you need to define a period for your project. Does your project focus on neighborhood change or the shrinking of Central Asia's Aral Sea? Then you need to establish a temporal extent. Determining the temporal period helps you determine your project's needed attributes. It also makes you aware that you may have to look for features that are no longer present on the landscape.

3. Determine the Attributes for Each Feature Type

As described in Chapter 1, attributes are the characteristics of features. You need to identify the required attributes for each feature type. The more you can do this before you collect your data, the less you will retrace your steps and collect additional attributes later. Again, look to the project's goals for some clues to the necessary attributes. Also consider how you will analyze the features. You cannot use some analytical processes (like many statistical tests) if the attribute values that you collect are in an improper form to be used in a particular analytical process (see levels of measurement below).

One other thing to consider at this point is that some attributes (like a polygon's area, a line's length, and even the number of point features falling within polygon features) can be generated automatically by the software. Additional attributes can be created by multiplying, dividing, adding, subtracting, truncating, or concatenating attributes with other attributes, numbers, or characters.

4. Determine How the Features and Their Attributes should be Coded

Once you have decided on the features and their attributes, determine how they will be coded in the GIS database. Remember from Chapter 1, there is not just one way to code features. Although roads are usually coded as lines, they do not have to be.

Decide whether to code each feature type as a point, line, or polygon. Then define the format and storage requirements for each of the feature's attributes. For instance, is the attribute going to be in characters

(string) or numbers? If they are going to be numbers, are they byte, integer, or real numbers? You will have to establish these database parameters before you enter data into the GIS. Look at the example below (Figure 2.4). Listed are some attributes (under Field Name) relating to the feature “streets”. Notice that street "LENGTH" has a data type called double (a type of real numbers), and in this case, the database will store up to 18 numbers including 5 decimal places for the length of each individual street.

Figure 2.4: Each feature’s attribute needs to be coded.

It is critical that you think about the value of your attributes before you code. Obviously, if one street segment needed room for 9 numbers to report its length, than 8 is not enough and the correct value could not be entered without modifying the field’s length.

Also, while thinking about your attribute values, consider where it fits on the "levels of measurement" scale

with its four different data values: nominal, ordinal, interval, and ratio. Stanley S. Stevens, an American psychologist, developed these categories in 1946. Although Steven's classification is widely used, it is not universally accepted. Some researchers have problems with the categories, and others with how categorization affects research. For our purposes, it is a useful way to conceptualize how data values differ, and it is an important reminder that only some types of variables can be used for certain mathematical operations and statistical tests, including many GIS functions. The different "levels" are depicted in Figure 2.5 and demonstrated using an example of a marathon race:

Nominal	Ordinal	Interval	Ratio
Runner ID	Order Finished	Time of Day Finished	Race Time
238	1	10:10	2:30
143	2	10:20	2:40
14	3	10:25	2:45
.	.	.	.
.	.	.	.
.	.	.	.
301	450	20:20	12:40

Figure 2.5: Levels of Measurement.

Nominal data use characters or numbers to establish identity or categories within a series. In a marathon race, the numbers pinned to the runners' jerseys are nominal numbers (first column in the figure above). They identify runners, but the numbers do not indicate the order or even a predicted race outcome. Besides races, telephone numbers are a good example. It signifies the unique identity of a telephone. The phone number 961-8224 is not more than 961-8049. Place names (and those of people) are nominal too. You may prefer the sound of one name, but they serve only to distinguish

themselves from each other. Nominal characters and numbers do not suggest a rank order or relative value; they identify and categorize. Nominal data are usually coded as character (string) data in a GIS database.

Ordinal datasets establish rank order. In the race, the order they finished (i.e. 1st, 2nd, and 3rd place) are measured on an ordinal scale (second column in Figure 2.5). While order is known, how much better one runner is than the other is not. The ranks 'high', 'medium', and 'low' are also ordinal. So while we know the rank order, we do not know the interval. Usually both numeric and character ordinal data are coded with characters because ordinal data cannot be added, subtracted, multiplied, or divided in a meaningful way. The middle value, the "median", in a string of ordinal values, however, is a good substitute for a mean (average) value.

The **Interval** scale, like we will discuss with ratio data, pertains only to numbers; there is no use of character data. With interval data the difference—the “interval”—between numbers is meaningful. Interval data, unlike ratio data, however, do not have a starting point at a true zero. Thus, while interval numbers can be added and subtracted, division and multiplication do not make mathematical sense. In the marathon race, the time of the day each runner finished is measured on an interval scale. If the runners finished at 10:10 a.m., 10:20 a.m. and 10:25 a.m., then the first runner finished 10 minutes before the second runner and the difference between the first two runners is twice that of the difference between the second and third place runners (see third column 3 Figure 2.5). The runner finishing at 10:10 a.m., however, did not finish twice as fast as the runner finishing at 20:20 (8:20 p.m.) did. A good non-race example is temperature. It makes sense to say that 20° C is 10° warmer than 10° C. Celsius temperatures (like Fahrenheit) are measured as interval data, but 20° C is not twice as warm as 10° C because 0° C is not the lack of temperature, it is an arbitrary point that conveys when water freezes. Returning to phone numbers, it does not make sense to say that 968-0244 is 62195 more than 961-8049, so they are not interval values.

Ratio is similar to interval. The difference is that ratio values have an absolute or natural zero point. In

our race, the first place runner finished in a time of 2 hours and 30 minutes, the second place runner in a time of 2 hours and 40 minutes, and the 450th place runner took 5 hours (see forth column in Figure 2.5). The 450th place finisher took over five times longer than the first place runner did (12.667 hrs / 2.5 hrs = 5.0668). With ratio data, it makes sense to say that a 100 lb woman weighs half as much as a 200 lb man, so weight in pounds is ratio. The zero point of weight is absolute. Addition, subtraction, multiplication, and division of ratio values make statistical sense.

5. Determine the Base Map Reference Features

What features are helpful to include? Add reference features that help people orient themselves within your study area even if you are not going to analyze these features. Major roads, rivers, and principal buildings are good examples of features that help orient map readers. These secondary features are often the easiest features to find on the Internet, and sometimes they are bundled with GIS software. In short, having these base-map features may not be important for analysis, but they are important for clarity.

6. Determine your Project's Projection, Coordinate System, and Datum

Before you collect or look for data, you should decide on which projection, coordinate system, and datum to use. These three terms, collectively termed “projection parameters”, are discussed in Chapter 3, but it is important that these parameters remain consistent throughout your layers. Consistency enables you to properly overlay your feature layers to produce maps and analyze feature relationships. Figure 2.6 shows an example of how a parcel layer is not properly overlaid upon street centerline and building layers due to small differences in parcel layer’s projection parameters.

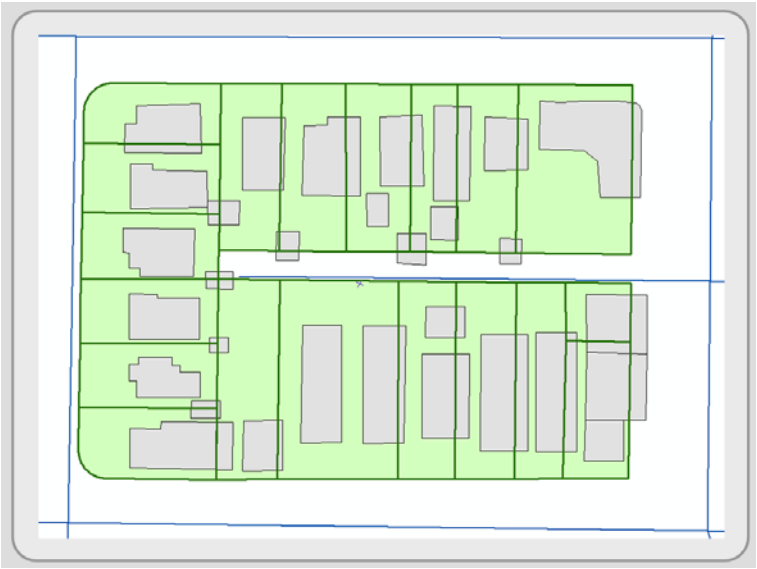

Figure 2 6: Due to differences in the projection parameters of the parcel layer, it is not lining up properly with the street centerline and buildings layers.

Deciding on the projection parameters upfront helps you select among nearly identical data files, if available, on the Internet because one file might already be in the desired projection, coordinate system, and datum. It is, however, rare to find everything the way you want it. Usually, you need to convert your layers, and GIS programs come with algorithms that convert between standard projections, coordinate systems, and datums. This topic is covered in Chapter 3.

Before choosing your projection parameters, do some investigating. If you are using mostly state GIS databases, it makes sense to use their parameters if they fit your needs. For instance, California has created its own Albers projection, and the vast majority of their GIS layers are in this projection. Local agencies, however, might use State Plane Coordinates or Universal Transverse Mercator. If substantial amounts of data are shared with a local jurisdiction, you may decide to use their projection parameters. You can change your projection parameters at any time in your project's life, but it takes time and organization.

The decision of which projection parameters to use can also be based on many other considerations. The project's scale, the desired projection properties (like the preservation of area, which is discussed in Chapter 3), the needed precision of positional accuracy, the attributes to be derived internally, and the

project's location are among those questions that need to be evaluated before selecting a projection. Know something about projection parameters before you choose them for your project. For instance, if you desire to calculate area from the shape of your polygons, you need to use an equal area projection. If you do not use an equal area projection, the area measurements will be inaccurate.

Before actively collecting and inputting data, consider the six items listed on page 15. Next, find out what GIS datasets already exist, and—for those datasets that do not exist—the different ways to create the needed data. The next two sections of this chapter deal with these topics.

One last suggestion, when planning a GIS project, talk with GIS technicians and analysts at government agencies or researchers that have put together similar GIS projects. Their suggestions will save you time, and they could become a great contact for troubleshooting problems and providing you with needed GIS datasets.

PHASE 2: DATA ACQUISITION

In the data acquisition phase, you obtain the data for your GIS. Getting all the data together (and in a suitable format) is the most costly and time-consuming task for any GIS project. Most estimates suggest that between 75 to 80 percent of your time is spent collecting, entering, cleaning, and converting data (phases 2 and 3 of this chapter). Before proceeding to data sources, we define the term “data” and look at various data issues including accuracy, precision, and metadata. Figure 2.7 features a list of data acquisition tasks.

Figure 2.7: Key tasks of the data acquisition phase.

Data

Data are frequently called facts. This definition suggests that they are pure, but all data are selected for a particular purpose and are shaped by that purpose. The term “raw data” implies an even greater purity—an objective truth—but even the most objective scientist has beliefs and knowledge that underlie data collection. Data obtained for any project are preconceived.

You can classify data as either primary or secondary and observable or non-observable. The geographer Frank Aldrich devised this simple data model (Figure 2.8) to depict the different data categories. Primary data (the light colored inner circle) are measurements that you or your team collects. They are usually derived from experiments or from fieldwork. Secondary data (the doughnut shaped darker ring) are datasets that someone else collects. These datasets, collected from experiments or fieldwork, were collected for a purpose other than your own. Most researchers prefer primary data because they have not been previously conceived and shaped. Still, secondary datasets are tremendously valuable if you determine how and why they were collected and if your project can accept those preconceptions.

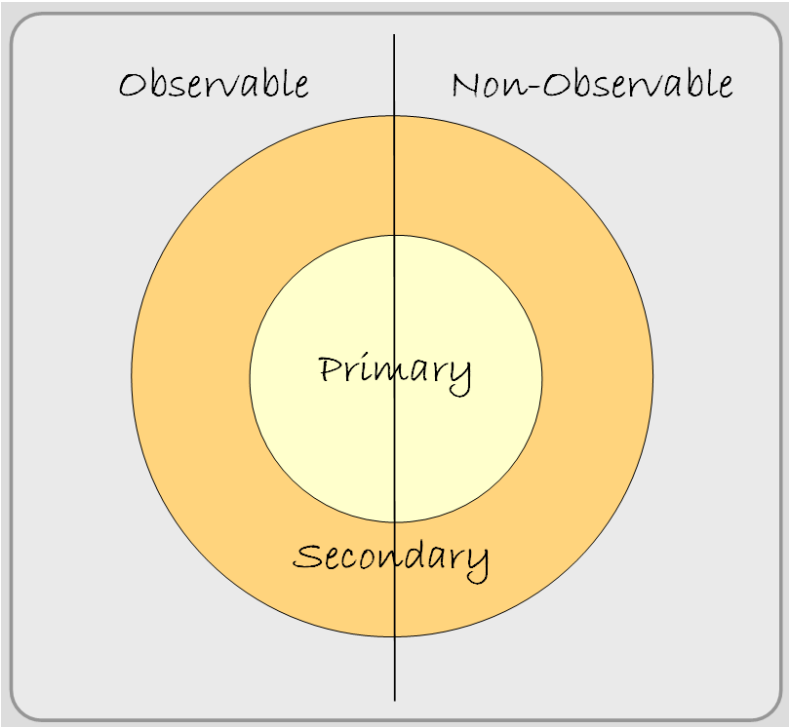

Figure 2.8: Aldrich's data model.

You can also classify data as either observable or non-observable. Notice in the figure above that a vertical line bisects the primary and secondary circles. Observable data (to the left of the bisecting axis) are when someone or something observes the characteristic or the behavior. Non-Observable data (to the right of the bisecting axis) are when respondents are asked questions in an interview or on a questionnaire, but the data gatherer does not physically observe the characteristic or behavior.

Combinations of these two types of characteristics are made. Primary Observable data are datasets you collect and observe. Secondary Observable data are datasets someone else collects and observes. Satellite imagery is an example. Primary Non-Observable data are datasets you collect, but the characteristic or behavior is not observed. Surveys and interviews that you administer are examples. Secondary Non-Observable data are datasets that someone else collects and does not observe. Census data falls into this category.

Data Evaluation

What is the quality of your data? Good datasets are both accurate and precise. There could be both spatial location problems and attribute errors. It is critical that you evaluate your data, especially your secondary datasets, because errors (if not corrected) can make a project's results worthless. Do not be lulled into a false sense of security when you receive or construct a dataset. Instead, evaluate your data. To better understand the types of errors that occur in GIS datasets, we turn to the term "error" and two parts of its definition.

Errors are obviously mistakes, and in the context of database errors, they are the result of two things: inaccuracies and data imprecision. Accuracy is the quality to which the data matches true and accepted values. In a layer, there can be inaccuracies in the location of features (and in their attributes). For example, features may be located in the wrong locations, not present when they should be, or a fictitious feature may be entered where one does not exist.

Precision is the exactness of a measurement or value, and it refers to the degree the value can be reproduced by similar data collection techniques. In other words, precision is a measure of exactness and repeatability. The precision of a feature's location could be accurate to a few inches or perhaps it has an acceptable precision of 10 meters. 10 meters may not seem precise, but it is usually good enough for many projects.

Many types of errors affect a GIS dataset's quality. Some are obvious, but others are difficult to uncover. Learn to recognize error and decide on an acceptable degree of precision for your project. The following discussion looks at the types of errors one might uncover. Chapters 3 and 4 address how to identify and fix these errors.

Accuracy Errors

1. How old is the dataset? This is one of the most important questions to have answered because some datasets may be too old to be useful. If it exists, a metadata file, a file that describes a feature layer, is a good place to start. Metadata files are discussed later in this chapter.

2. The dataset may have locational inaccuracies. As described above, features may be sloppily positioned, omitted, or placed where they do not exist. This is usually the result of faulty fieldwork and observations, input error, or problems associated with converting data. On this last point, the accurate conversion of hard-copy maps to digital form is a major challenge because processing errors occur during digitizing and scanning (defined later in this chapter).
3. The attributes may be faulty due to keyboard error, faulty observation techniques, defective or non-calibrated instruments, or researcher bias.
4. If you are using secondary data, discover how the digital dataset was created and whether it underwent any digital conversion. Conversion problems occur when processing data from different formats, projections, data models, and resolutions. There can be inconsistencies (even slight) during the translation that alters geographic position and (in the case of raster resampling) the cell values.

Precision Errors

1. The dataset may have positional imprecision due to the scale of the original maps. A map with a scale of 1:24,000 illustrates finer details than a smaller scale map of 1:500,000. With a vector-based GIS, as you zoom into your map, the feature locations may look precise, but its precision is based on the scale of the original map. It might not be as precise as it looks. Using a raster GIS, you can generate imprecision by starting with a low-resolution layer (large pixel size) and increasing its resolution. The pixels are capable of holding a higher resolution of data, but the precision will not improve.
2. Pixel resolution can also contribute to a dataset's positional imprecision. As discussed in Chapter 1, if a layer has a large pixel size, it may poorly represent the positional precision of its features. Small and narrow features will swim within large pixels. Their precise placement is unknown.
3. Positional imprecision can also occur if feature position is difficult to determine. Some features like streets and parcels are fairly easy to place, but there are features like soils, vegetation, and climate regimes that have fuzzy or less discrete borders. Some lines across a map are value judgments. The data may be accurate; they are just not precise.

Metadata

Given these possible errors, you can understand the danger and uncertainty of using undocumented data. Metadata is a data quality document, and its frequently repeated definition is “data about data” (although it is perhaps more accurate to define it as "information about data"). It describes the attributes and the location of the features in the layer. It gives you an impression about the dataset's accuracy and precision. Metadata includes basic information about the dataset, including a description and if there are any use rules. Good metadata files should provide the answers to the following questions:

- What is the dataset's age?
- What is the area covered by the dataset?
- Who created it?
- How was it constructed (digitizing, scanning, overlay, etc.)?
- What projection, coordinate system, and datum are used?
- What was the original map's scale (if applicable)?
- How accurate and precise are the locations and attributes?
- What data model (vector or raster) does the layer use?
- How were the data checked (both location and attributes)?
- Why were the data compiled? What was their need or motivation?

After looking over the metadata file, you should ask yourself and your colleagues an additional question: is the data provider reliable? The presence and condition of the metadata might help answer this question.

Also within the metadata are data dictionaries that describe all of the features’ attributes. As you saw in Figure 2.4, attributes frequently have odd, short field names, like Zblack_00, that are difficult to decode. Metadata describes the attributes more fully. In addition, the values that go into these attribute fields are often coded with short values (or abbreviations) instead of using longer words, which are more subject to keyboard error. Data dictionaries decipher these codes and abbreviations. The metadata file in Figure 2.9 is an example. It documents a California weather station GIS layer from the California Spatial Information

Library (CaSIL). It has been modified (shortened) for the purposes of this e-text.

Metadata is important. Years ago when GIS was in its infancy, there were no standards relating to data documentation and very little metadata existed. GIS layers were created with only a particular project's specifications in mind, and many of the details went missing when the personnel that created the dataset moved on. As GIS datasets became numerous and agencies began sharing data, a common set of specifications arose to describe the GIS layers. A metadata file is generally attached or closely associated with a GIS layer.

LIBRARY: COUNTY, CA

LAYER NAME: WEATHERSTATIONS

COVERAGE NAME: WXSTNA

COVERAGE DESCRIPTION: The 'WEATHERSTATIONS' layer represents U.S. Weather Service surface weather observing stations. Many of the stations are located at airports. Others are located in or near USWS offices. This coverage was created from the file STATIONS.ZIP available from vmd.cso.uiuc.edu (anonymous ftp).

VITAL STATISTICS:

Datum:	NAD 27
Projection:	Albers
Units:	Meters
1st Std. Parallel:	34 00 00 (34.0 degrees N)
2nd Std. Parallel:	40 30 00 (40.5 degrees N)
Longitude of Origin:	-120 00 00 (120.0 degrees W)
Latitude of Origin:	00 00 00 (0.0 degrees)
False Easting (X shift):	0
False Northing (Y shift): -4,000,000	
Source:	vmd.cso.uiuc.edu (University of Illinois)
Source Media:	ftp file
Source Units:	degrees/minutes
Source Scale:	unknown
Capture Method:	Converted to ARC/INFO using generate
Conversion Software:	ARC/Info Rev 6.1.1
Data Structure:	Vector
ARC/INFO Cov. Type:	Point
ARC/INFO Precision:	Single
ARC/INFO Tolerances:	default
Number of Features:	125
Layer Size:	.22 MB
Data Updated:	January 1994

DATA DICTIONARY:
DATAFILE NAME: WXSTNA.PAT; RECORD LENGTH: 51

COLUMN	ITEM NAME	WIDTH	OUTPUT	TYPE	N.DEC
1	AREA	4	12	F	3
5	PERIMETER	4	12	F	3
9	WXSTNA#	4	5	B	-
13	WXSTNA-ID	4	5	B	-
17	X-COORD	4	12	F	3
21	Y-COORD	4	12	F	3
25	NAME	16	16	C	-
49	ABB	3	3	C	-

AREA: The area of the polygon in square coverage units.
PERIMETER: The length of the polygon perimeter of the polygon in coverage units.
WXSTNA# : The software-assigned unique integer identification number.
WXSTNA-ID : A user-assigned identification number.
X-COORD : The Arc/INFO supplied x value for longitude in lat/long.
Y-COORD : The Arc/INFO supplied y value for latitude in lat/long.
NAME : The NWS-supplied station name.
ABB : The 3-character NWS station identifier.

DATA QUALITY ASSESSMENT:
The following are subjective comments regarding this data. Feature completeness is unknown. Feature accuracy is only to 1/100th of a degree of latitude or longitude. Attribute accuracy and completeness are also unknown.

DATA CONTACT:
Contact Name: ABCD; Contact's Phone: 916-### ####
DOCUMENTATION DATES: 10/27/####

Figure 2.9: A typical metadata file that provides information on both spatial and attribute data.

Despite its importance, metadata is still neglected. It is time consuming, but it is essential in an environment where data sets are shared and the background of the data must be known.

On-Line GIS Data

The Internet is a great place to start looking for data. If you find existing GIS datasets that serve your purpose and passes your specifications, it saves you time and money. A search may reveal multiple copies of what seems to be the same data, but check the details—examine the metadata—because minor differences might make one dataset better than the other. Much base map data (countries, states, counties, major roads, rivers, township and range) exists on the Internet.

It would be convenient to retrieve all of your GIS datasets from the Internet, and although more and more data are available, the Internet will not provide you with everything you need. When searching the Internet, don't just look for GIS files. Search for data that might be GIS compatible. Perhaps a spreadsheet can be modified and linked to an existing geographic layer. CAD files, aerial photographs, and satellite imagery might also be relevant. Finding good digital datasets frees you from collecting and entering the data yourself.

Other Sources of GIS Data

Most likely, you will contact GIS personnel at various government agencies, commercial data sources, and organizations about their data sources. Ask questions about the datasets accuracy and completeness. If you think their datasets might be helpful to you, ask them for permission to obtain and use the data. Frequently, a short agreement will be written and signed that allows one to use the data under certain conditions.

Many data companies modify “public” data to create a “value-added” product that you can purchase and load directly into your GIS. “Value-added” datasets usually originate from a government agency or an organization that creates the basic GIS dataset, but a commercial company obtains the data and "improves" it by adding attributes or improving its spatial precision. The commercial company can then sell the "value-added" portion of the data. Many of these datasets can also be obtained over the Internet.

Converting from one GIS format to another.

When obtaining GIS data from the Internet or from other sources, it requires extensive preprocessing to make it work with your other GIS datasets. Initially, the newly acquired dataset may require "extracting" (unzipping). They are often compressed to make them smaller for storage on a CD or to make their downloading quicker. If the acquired file's type is ZIP, TAR or GZ, it needs to be extracted, which can be done with free or cheap software (like WinZip). At times, extracting files can be more complicated if the compressed files are nested within a compressed file. This requires extracting the first compressed file and then extracting the nested files.

GIS datasets are typically stored in one of the leading GIS software formats (frequently as a shapefile) or in a format specified by the U.S. federal government. Fortunately, most GIS packages read or convert many of the most common GIS formats. At times, however, you may need access to "third party" software to read the data and export it into a format that your GIS program reads.

Even after the datasets are extracted and converted into your GIS format, they will require further processing. Many types of data preprocessing, manipulation, and conversion are discussed in Chapters 3 and 4.

PHASE 3: DATA CAPTURE

Introduction

When you have exhausted your contacts and the Internet, it is time to capture the data yourself. In this phase, you create new GIS datasets from both digital data that are not currently in a GIS format and from non-digital, hard-copy data sources. Examples of digital and non-digital data sources include maps (hard-copy and digital), aerial photographs (hard-copy and digital), questionnaires, field observations, digital satellite imagery, survey data, and Global Positioning System (GPS) coordinates.

The data capturing phase is often tedious, laborious, and frustrating, but necessary. Figure 2.10 breaks

the stage into three main steps, which are described in the text below:

Figure 2.1: The key steps in data capture phase. Here you digitize hard-copy maps and data directly into your GIS or transform existing digital data into a format your GIS reads.

Converting Digital Data

This section looks at digital datasets that are currently not in a GIS format, but that are often manipulated to create GIS layers. These sources include automated surveying, photogrammetry, GPS, Light Detection and Ranging (LIDAR), non-GIS mapping programs like Computer Aided Design (CAD), and remote sensing imagery. This chapter emphasizes the conversion of GPS data and remote sensing imagery. Importing spreadsheet data, a frequent way of importing attribute data, into a GIS is covered in Chapter 4.

Automated surveying uses electronic data capturing instruments like theodolites, electronic distance measurement (EDM) systems, and total stations to capture spatial and attribute data. The most sophisticated of these instruments is the total station that combines the theodolite's angle-measuring capabilities with the EDM's distance calculations. Surveyors download the distance and direction data from their instruments directly into many vector-based GIS programs. The data, however, usually requires pre-processing before it can be used to make a map.

Figure 2.2: Total station, a surveying instrument, measures distance, direction, and location. Image comes from Illinois State University, Geography Dept.

Photogrammetry obtains accurate measurements from aerial photographs. Photogrammetric techniques determine ground distances and directions, heights of features, and terrain elevations. Photogrammetry creates GIS data through 3-D stereo digitizing and by producing spatially rectified aerial photographs that can be entered into the GIS as a layer.

GPS (Global Positioning System) is a radio-based navigation system that uses GPS receivers to compute accurate locations on the Earth's surface from a series of orbiting satellites. With a small, inexpensive, hand-held GPS receiver (Figure 2.12), you can determine your location usually within about three meters. Using two GPS receivers simultaneously (called Differential GPS or DGPS) or using a Wide Area Augmentation System (WAAS)-enabled GPS receiver, which uses satellites and ground stations that provide GPS signal corrections, you can get sub-meter accuracy.

Figure 2.3: Typical low-end GPS device.

The U.S. GPS network, called NAVSTAR (Navigation with Satellite Timing and Ranging), has at least 24 satellites that orbit in six planes around the Earth (see Figure 2.13). The network's configuration secures at least four satellites—the minimum number of satellites needed to capture location data—above the horizon for every point on Earth.

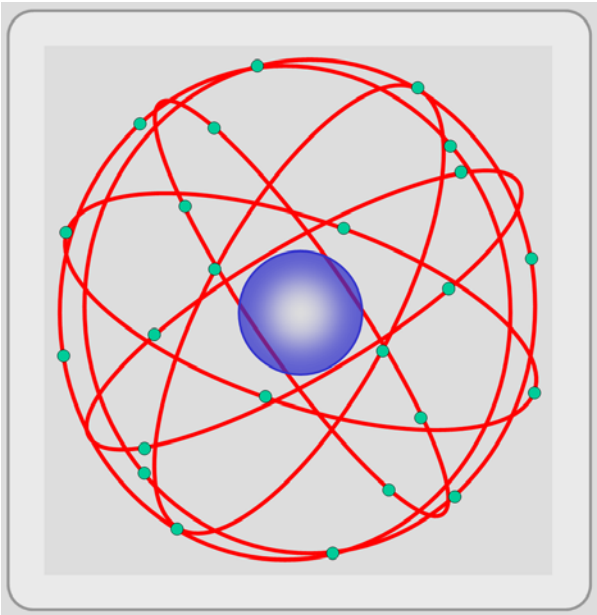

Figure 2.4: NAVSTAR GPS constellation of satellites.

The U.S. Department of Defense (DoD) developed and controls NAVSTAR, and DoD can turn the whole system off, as they briefly did immediately following the terrorist attacks of September 11, 2001. DoD monitors and tracks the satellites (which are equipped with radio transmitter/receivers and a set of atomic clocks) from five stations across the globe where they compute precise satellite orbital and clock corrections. These corrections are transmitted from the master control station at Schriever Air Force Base in Colorado to the satellites, which make the adjustments. All of the satellite's locations are precisely known, and by knowing their exact locations, we can determine the location of every point on Earth with a GPS receiver.

This is possible because each satellite transmits a unique radio signal, which can be received by GPS receivers. Using this signal, your GPS receiver calculates the distance to each of the four satellites it is tracking by the amount of time it takes for the signals to travel from the satellites to your receiver. This is a high-tech version of triangulation (see Figure 2.14), called trilateration. The first satellite locates you somewhere on a sphere (top left of Figure 2.14). The second satellite narrows your location to a circle created by the intersection of the two satellite spheres (top right). The third satellite reduces the choice to two possible points (bottom left). Finally, the forth satellite helps calculate a timing and

location correction and selects one of the remaining two points as your position (bottom right).

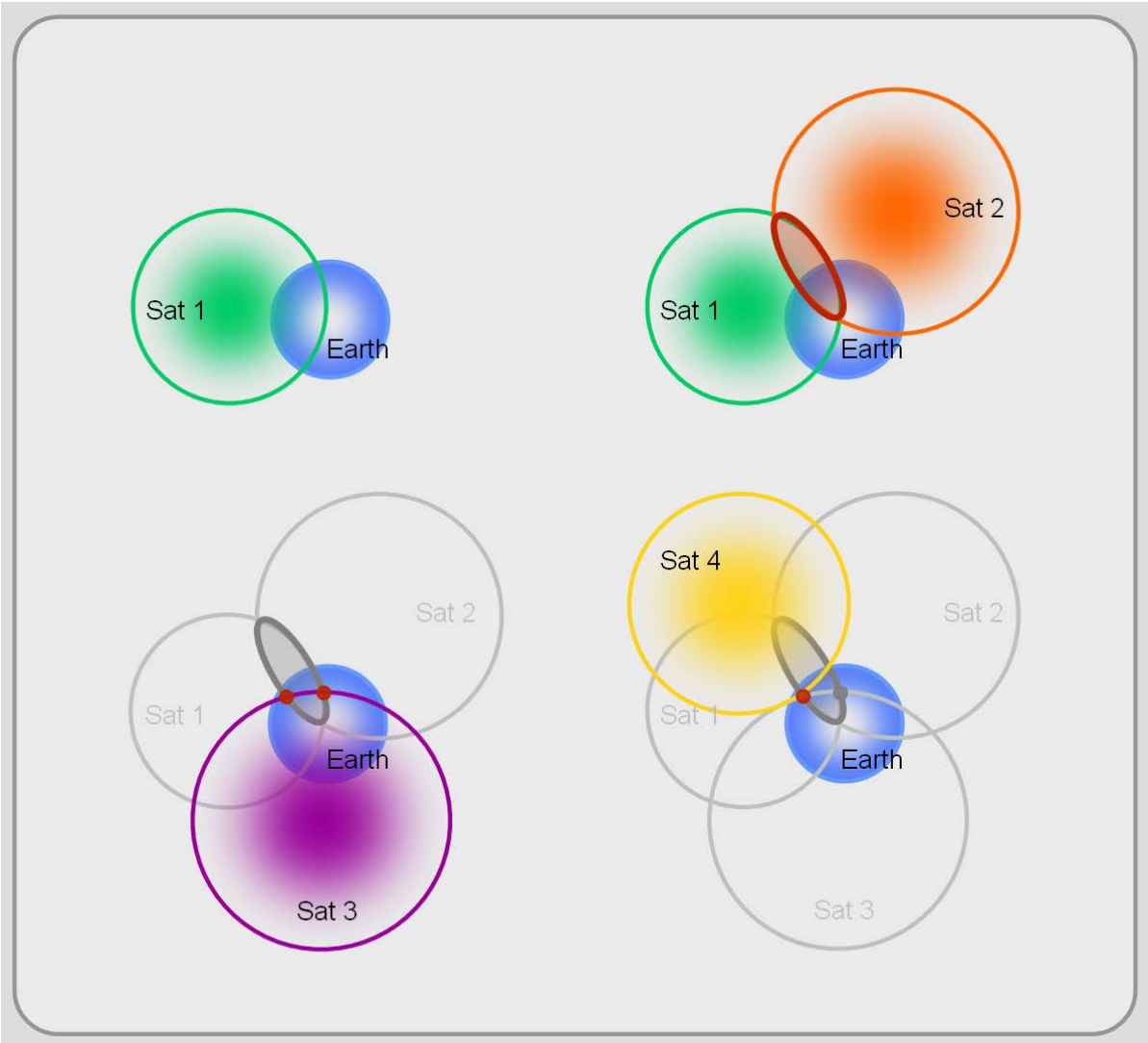

Figure 2.14: GPS satellite trilateration.

When dealing with satellites thousands of miles from the Earth's surface, GPS's accuracy is amazing. Still, the technique is not without error. Small errors in the receiver's clock, variations in the satellite's orbit, atmospheric conditions that slow radio waves, and radio signals that bounce (called “multipath” or "ghosting") off buildings and cliffs are some possible distortions. In addition, GPS has difficulty penetrating thick forests and urban canyons created by tall buildings.

Another source of error is Geometric Dilution of Precision (GDOP). This is the spatial relationship between the GPS receiver and each of the potential satellites. In general, the fewer the satellites

available, and the closer they are clustered, the less accurate your readings. GPS receivers try to avoid GDOP by selecting the set of satellites that provide the least error. It chooses satellites that are well above the horizon, minimizing atmospheric thickness and interference from buildings, yet not so high that they are clustered together.

GPS is a major data input tool for GIS. Most receivers, even inexpensive units, contain a hard-drive where you can log your positions. Each logged position is called a waypoint, and together, waypoints depict the location and extent of the features you enter in the field. They can be downloaded from GPS receivers (sometimes with the help of a separate software program) and imported into many GIS programs.

Some of the more expensive GPS units have “feature lists” that streamline the data entry process. Feature lists are databases that you define to contain a list of the possible features you will locate. These lists have associated attributes for each feature type, and drop down lists of common attribute values can also be predetermined for each attribute to save time in the field. Both the feature locations and their associated attributes can be downloaded into your GIS.

LIDAR (Light Detection and Ranging) is a remote sensing technology that uses laser light pulses to measure the distance to a surface. It is similar to other types of radar, but uses light instead of radio waves. Airborne LIDAR systems have resulted in topographic layers that depict the tops of ground-based features better than traditional remote sensing and radar methods, and this results in topographic layers that portray the shape of our cities (including the widths and heights of buildings) and forest canopies more accurately. Figure 2.15 is an example of LIDAR image in an urban area.

Figure 2.5: LIDAR image depicting feature heights.

Both vector and raster data models use LIDAR data. Many GIS programs have automated routines that convert the x, y, and z (elevation or altitude) points first into vector point files, which can be converted into raster layers (rasterization is described in Chapter 3) where additional processing and analysis can take place. In vector systems, LIDAR points frequently generate topographic layers called Triangular Irregular Networks (TIN), which can be used to represent terrain in vector-based systems.

Conversion of Spatial Data from Non-GIS Programs is a common way to acquire data. Other than GIS, there are additional computer programs that display, create, and edit vector spatial data. A brief list includes Computer Aided Design (CAD) systems (like AutoCAD, MicroStation, and ArchiCAD) and vector drawing programs (like Adobe Illustrator, Corel Draw, and OpenOfficeDraw). The conversion process varies based on the format of the dataset and the import/export capabilities of both the host

software and the GIS program. Capturing and importing CAD and drawing data into vector-based GIS often involves converting the digital data from the host format into the desired GIS format or into a format that can be easily read by your GIS. Sometimes the conversion process is that easy. Sometimes it's not; the data may need to be entered into a "third party" format that both programs read or export to. AutoCAD files are an example. They usually need to be saved in DXF format within the CAD software before many GIS programs can read them.

Remote Sensing imagery is an important source of raster data. This section briefly describes satellite remote sensing, digital image processing, and the conversion of remotely-sensed imagery into GIS.

The word "remote" means "from a distance." Sensing, in this case, means "to record." So a basic definition of remote sensing is to record from a distance.

To understand remote sensing, we need to understand some important core concepts. First, all features on the Earth's surface reflect and absorb the sun's radiant energy. They also emit and reflect back parts of that energy. The amount and type of radiation that is emitted and reflected from the Earth's features depends on the properties of those features (see Figure 2.16).

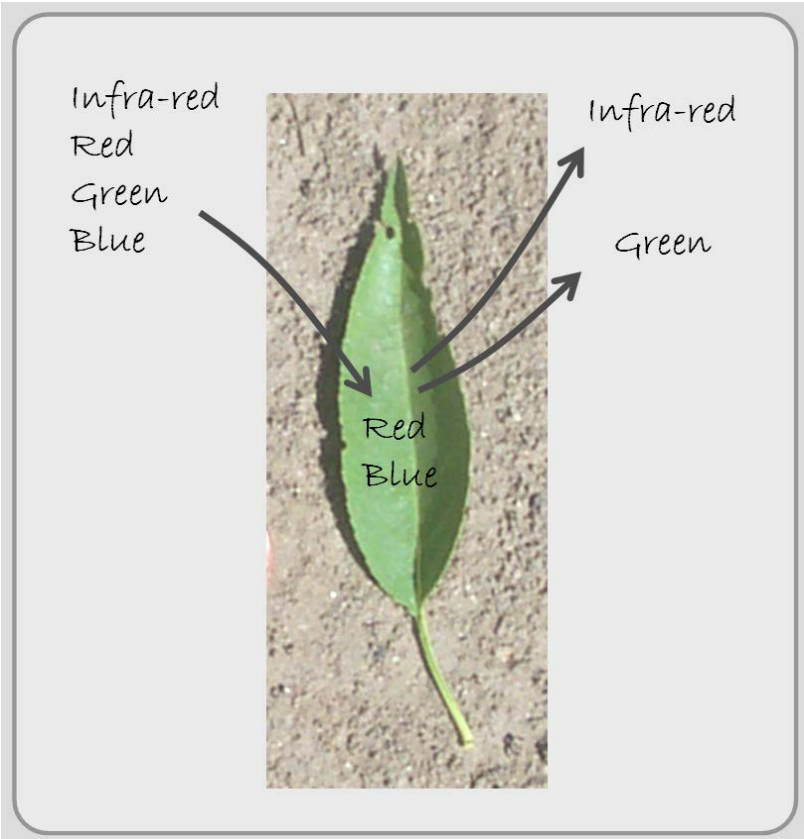

Figure 2.16: Bands of light including infra-red, red, green, and blue are emitted from the sun. The green leaf absorbs the red and blue bands, but largely reflects the infra-red and green bands thus appearing to our eyes green.

What our eyes see is the reflected radiation within a very small part of the electromagnetic spectrum (see Figure 2.17). We see the blue, green, and red bands of light, known as the visible spectrum, but other wavelengths are present that we do not sense. Satellites, however, have sensors that record not only the visible spectrum, but also infrared, near infrared, and thermal infrared bands. Satellite sensors, however, vary by maker, and so they record different wavelengths. The precise bands (wavelengths) a satellite captures is referred to as spectral resolution.

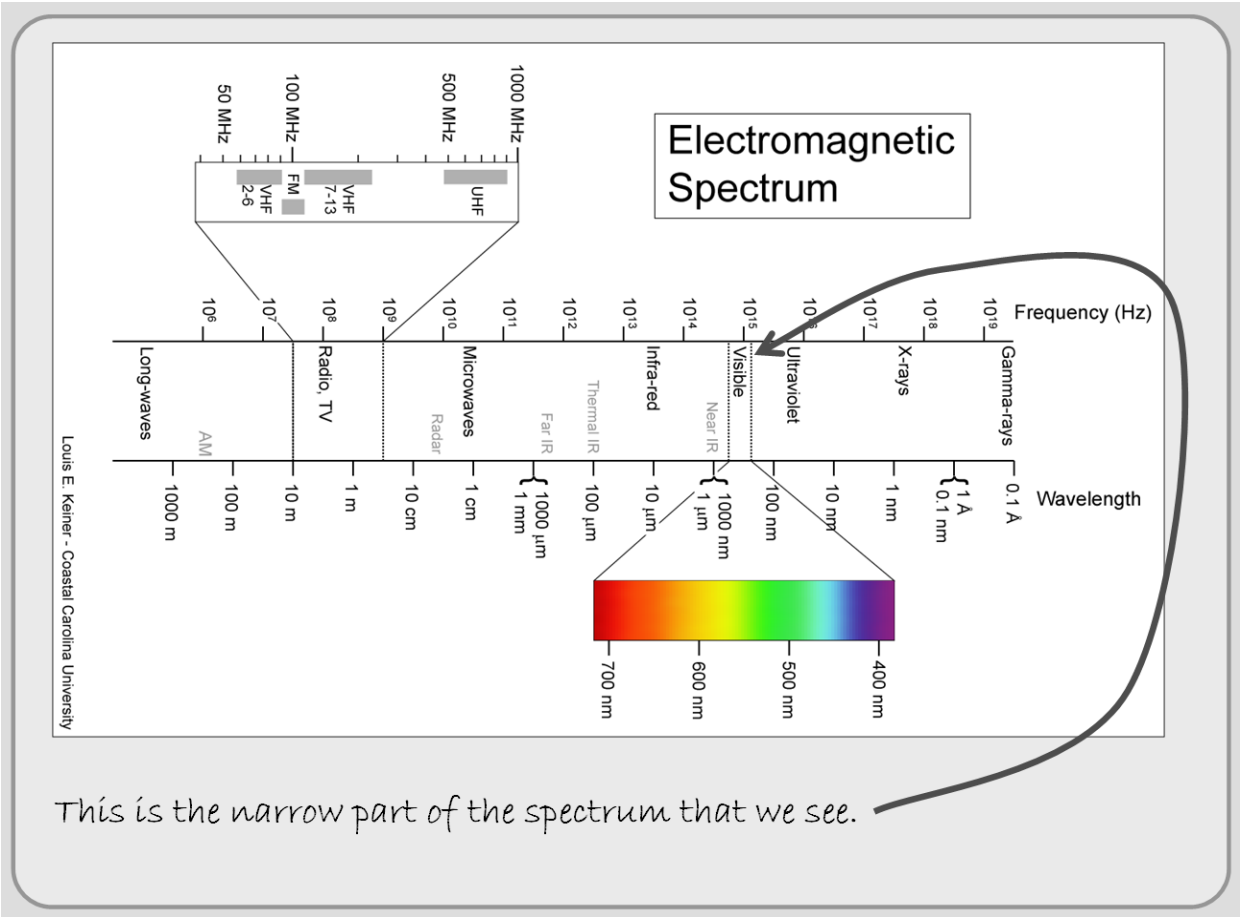

Figure 2.17: The electromagnetic spectrum with the visible waves highlighted. Primary image is by Louis Keiner, Coastal Carolina University. See licensing at http://commons.wikimedia.org/wiki/File:Electromagnetic-Spectrum.png.

Through telemetry, satellites transmit each of these individual wavelengths into remotely sensed data files, called images, which measure the reflected and emitted electromagnetic energy from the Earth's features. The images are raw meaning they are unprocessed, and they need to be digitally enhanced and combined to highlight particular features (land use, climate changes, agricultural productivity, and environmental properties).

Like any raster image, remote sensing images consist of pixels. Each pixel in a data file records a particular wavelength (band of light) of the electromagnetic spectrum for a specific chunk of the Earth's surface. The pixel size is its spatial resolution, and it differs from spectral resolution in that it involves pixel size rather than a particular wavelength or band of light recorded by the satellite. Like spectral resolution, different satellites acquire data at different spatial resolutions.

Different types of features emit and reflect back different intensities of each wavelength of light, so remote sensing images are combined and manipulated to highlight specific features. This procedure of combining images to isolate features is called digital image processing. Picking the proper images comes with experience and knowledge, but it is based on the feature's spectral signature, which depicts the percent of each band of light that is emitted and reflected back off the features. Each spectral signature is kind of a feature's unique fingerprint.

Digital image processing is done in one of many specialized remote sensing software packages like Erdas Imagine or ER Mapper (but some raster-based GIS programs, like Idrisi, are capable of many image-processing functions). Once processed, the images can be entered into your GIS. Some GIS programs read some of the major remote sensing image formats, but you may need to export the image to a format that your GIS program reads. The processed images are usually used in raster-based GIS as a continuous data surface, such as climate data. The images can be used in vector-based systems simply as a reference background image or to trace vector features over (a process called head's up digitizing, see next section).

Converting Non-Digital Data

Existing, hard-copy maps and aerial photographs (physical paper documents) are a major source of spatial data for GIS. Different processes, including digitizing, scanning, and "heads up" digitizing, exist to input these hard-copy sources into GIS. For most projects, you will need to capture both the spatial location of the feature and some of its attributes.

Data input is usually the major bottleneck in the development of a GIS database, and converting hard-copy data from maps, aerial photographs, printed reports, and field notebooks is often the least desirable option because it is tedious and time consuming. Still, it is a way of making sure that you get a certain level of accuracy and precision for your project.

Scanning is a popular way to convert hard-copy maps and aerial photographs into digital images. The resultant scanned image is a raster file, arranged as an array of pixels in columns and rows. Scanners capture what is on the original document by assigning a color or grayscale value to each pixel in the array. The pixel value is based on the intensity of the color or gray shade on the original document.

Scanners come in a range of types, sizes, and degrees of sophistication. Scanner types include flat bed, sheet fed, drum, and video. Flat bed (or desktop) scanners are the most common and consist of a glass board where you lay the documents you want to scan. It works like a copier, but the output saves to an image. Usually, they scan areas no larger than a legal-sized document but a couple brands make flat beds up to 24” x 36".

For hard-copy documents larger than 24” x 36”, you can choose a sheet fed (see Figure 2.18), drum or video scanner. Sheet fed scanners work like a fax machine. The document moves past a scanning head, and the image is captured. It scans only loose sheets and the image is sometimes distorted if the document is not handled properly. A few varieties can handle rigid maps on poster board. Drum scanners are a great alternative due to their size and accuracy, but they are expensive, which generally limits their use to large commercial firms and government agencies. You feed maps into them and its rotating drum systematically scans the data. Video scanners use a high-resolution camera system to scan—incrementally—portions of the map. They have become an alternative to drum scanners because they are cheaper and their accuracy is improving (still, there are geometrical distortions and uneven brightness values across their scanned images). They are, however, extremely fast; most scans from a video scanner take about a second.

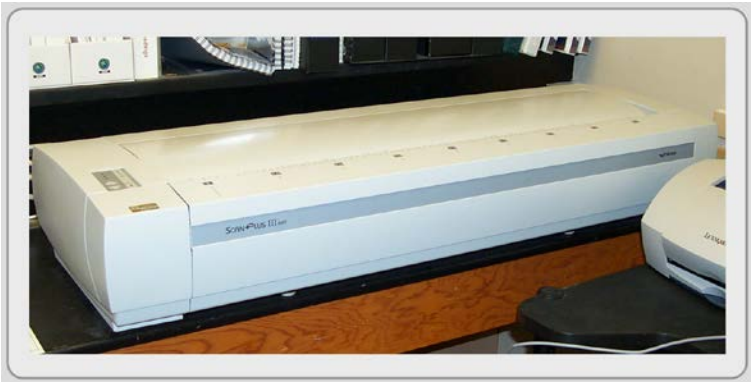

Figure 2.18: Wide-format, sheet-fed scanner.

For many GIS applications, you may also need to consider the scanner's resolution and its bit depth. Resolution is the detail the scanners read. Scanners support resolutions from 72 to 1200 dpi (dots per inch), but most scanning for GIS applications is done at resolutions between 200 and 600 dpi. Higher resolutions create higher quality images, but they also create substantially larger file sizes. Also, consider that at the same resolution, color images are larger than grayscale images, which in turn are larger than black and white images.

Bit depth is the number of bits used to represent each pixel. The greater the scanner's bit depth, the more colors and shades of gray that can be represented in your digital image. For example, a 24-bit color scanner can represent 2 to the 24th power or 16.7 million colors. Greater bit depth, however, creates larger files.

Before scanning any document, remember to clean it—smooth out folds and tape any tears. This reduces spatial inaccuracies. Also, erase any marks you do not want to appear on the resultant image. Similarly, make wanted notes and highlighted features legibly on the hard-copy document if you want to see them on the scanned image.

Once scanned, remember that you have essentially a "dumb" image; its pixels do not have meaningful values (just a gray or color value taken from the scanned document). This limits its use because it is

just a pretty picture. Although future scanning technologies focus on generating vector features directly from scanned images (called auto-vectorization), these technologies are not ready. Scanning, however, remains beneficial because of a second input process called “heads up” digitizing.

Heads Up Digitizing

After you create a scanned image, you georeference it (see Chapter 3) and use it as a background image within your vector system. Then with the image at its proper geographic location, trace the features that appear on the scanned image. This process, called "heads-up" digitizing (or on-screen digitizing), is like manual digitizing (described below) but without a physical digitizing board. Instead, you see on the screen a scanned image in its correct geographic position, and, with your mouse, you trace the position of features into new or existing point, line, and polygon layers. In the example below, building footprint polygons are traced over an image to create a vector polygon layer of buildings.

Figure 2.19: Head’s up digitizing. By eye, trace the features on your georeferenced image.

After you trace each feature, enter the feature’s key identifier into the layer’s attribute table (see Figure

2.20). How you enter key identifiers depends on whether the feature's layer is brand new or whether you are updating an existing feature layer. If you are updating an existing layer, the field will already be present, and you simply need to fill in the feature's value in the appropriate field. If you are creating a new layer, you must define at least the key identifier's attribute field before digitizing. This is important because each feature needs to be identified uniquely. The rest of the attributes are usually created in a separate file (perhaps in Access or Excel) and "joined" to the feature's minimal attribute table after digitizing is complete. The joining process is covered in the second half of Chapter 4.

Figure 2.20: Entering key (unique) identifiers into the new layer's attribute table.

The procedure for digitizing features is similar to that used in "manual" digitizing (see below). Due to the increased popularity of heads-up digitizing, manual digitizing is less frequently used, but for certain applications it is an important way to input map data.

Digitizing involves tracing by hand the extent of features directly from a hard-copy map or photograph that is mounted onto a digitizer (see Figure 2.21), a large table or board with an imbedded electronic grid that senses the position of a pointer called a puck (a mouse like device).

Figure 2.21: Digitizer with puck.

All GIS packages have a specific procedure for manual digitizing. Generally, it involves three steps: mounting the map on the digitizer, establishing control points, and adding map features.

In the first step, you mount the map in the middle of the digitizer's surface with its corners held by masking tape. Stretch out any map creases, but try to obtain maps that are in good shape because spatial errors may result. If you are working with the same map over a series of days, you need to remount the map periodically because humidity can expand paper maps, and they can sag slightly.

Step two involves establishing your map's control points. Although software packages vary, establishing control points involves a routine where you identify at least four points that are common to both the mounted map on the digitizer and the map on your screen. To minimize error, pick control points that surround—and one that is within—the area to be digitized. Then point to the four control points on the digitizer's map with the puck, and in the same order, identify the same four locations on

the screen's map with the mouse. This results in a spatial connection between the map on the digitizer and the screen's map.

Once this connection is established, you can trace the map features on the digitizer (Step 3). Take the puck and position it directly over one of the feature's vertices (corners). Various buttons on the puck allow you to create vertices, delete vertices, and close polygons. Like "heads up" digitizing, you will want to enter at least the feature's key identifier as you digitize

GPS digitizing involves using a GPS receiver to record feature data in the field. Using GPS for point locations (waypoints) was described above, but mapping-grade GPS units, like Trimble's GeoXT, are capable of recording the nodes and locations of points, lines and polygons by following the feature's extent and registering a waypoint at each of its vertices.

Pilot Project

A pilot project is a rehearsal. Here you collect a small subset of the GIS datasets you require for the larger project. Then you input the data into the GIS, preprocess the datasets, analyze them, and create some output. When something goes wrong, you tweak the project's parameters until the process works smoothly.

Pilot projects give you the opportunity to "ground truth" your secondary data. Remember, it is foolish to believe these datasets are without flaws. You need to ground truth your GIS data to ensure that the datasets are representative of what's on the ground. It is done by traveling to your study area and using your eyes to verify your datasets. Normally, you do not need to generate statistics nor check every feature. Checking an informal sample is good enough. Likewise, if any of your datasets are old or some features have oddly shaped features, this is a good opportunity to check them too. If ground truthing is difficult due to access or the aggregation of attributes (as in census data), cross validate your GIS data with other sources like aerial photography, satellite imagery, assessor data, or demographic figures.