Chapter 3

Data, Information, and Where to Find Them

Maps are shared, available, and distributed unlike at any other time in history. What's more is that the process of mapping has also been decentralized and democratized so that many more people not only have access to maps but also are enabled and empowered to create their own maps. This democratization of maps and mapping is in large part attributable to a shift to digital map production and consumption. Unlike analog or hardcopy maps that are static or fixed once they are printed onto paper, digital maps are highly changeable, exchangeable, and as noted in Chapter 2 "Map Anatomy", dynamic in terms of scale, form, and content.

To understand digital maps and mapping, it is necessary to put them into the context of computing and information technology. First, this chapter provides an introduction to the building blocks of digital maps and geographic information systems (GISs), with particular emphasis placed upon how data and information are stored as files on a computer. Second, key issues and considerations as they relate to data acquisition and data standards are presented. The chapter concludes with a discussion of where data for use with a GIS can be found. This chapter serves as the bridge between the conceptual materials presented in Chapter 1 "Introduction" and Chapter 2 "Map Anatomy" and the chapters that follow, which contain more formal discussions about the use and application of a GIS.

3.1 Data and Information

LEARNING OBJECTIVE

1. The objective of this section is to define and describe data and information and how it is organized into files for use in a computing and geographic information system (GIS) environment.

To understand how we get from analog to digital maps, let's begin with the building blocks and foundations of the geographic information system (GIS)—namely, **data**[1] and **information**[2]. As already noted on several occasions, GIS stores, edits, processes, and presents data and information. But what exactly is data? And what exactly is information? For many, the terms "data" and "information" refer to the same thing. For our purposes, it is useful to make a distinction between the two. Generally, **data** refer to facts, measurements, characteristics, or traits of an object of interest. For you grammar sticklers out there, note that "data" is the plural form of "datum." For example, we can collect all kinds of data about all kinds of things, like the length of rainbow trout in a Colorado stream, the number of vegetarians in Alaska, the diameter of mahogany tree trunks in the Brazilian rainforest, student scores on the last GIS midterm, the altitude of mountain peaks in Nepal, the depth of snow in the Austrian Alps, or the number of people who use public transportation to get to work in London.

Once data are put into context, used to answer questions, situated within analytical frameworks, or used to obtain insights, they become **information**. For our purposes, **information** simply refers to the knowledge of value obtained through the collection, interpretation, and/or analysis of data. Though a computer is not necessary to collect, record, manipulate, process, or visualize data, or to process it into information, information technology can be of great help. For instance, computers can automate repetitive tasks, store data efficiently in terms of space and cost, and provide a range of tools for analyzing data from spreadsheets to GISs, of course. What's more is the fact that the incredible amount of data collected each and every day by satellites, grocery store product scanners, traffic sensors, temperature gauges, and your mobile phone carrier, to name just a few, would not be possible without the aid and innovation of information technology.

Since this is a text about GISs, it is useful to also define **geographic** data. Like generic data, **geographic** or **spatial data**[3] refer to geographic facts, measurements, or characteristics of an object that permit us to define its location on the surface of

1. Facts, measurements, and characteristics of something of interest.
2. Knowledge and insights that are acquired through the analysis of data.
3. Data that describe the geographic and spatial aspects of phenomena.

the earth. Such data include but are not restricted to the latitude and longitude coordinates of points of interest, street addresses, postal codes, political boundaries, and even the names of places of interest. It is also important to note and reemphasize the difference between geographic data and **attribute data**[4], which was discussed in Chapter 2 "Map Anatomy". Where geographic data are concerned with defining the location of an object of interest, attribute data are concerned with its nongeographic traits and characteristics.

To illustrate the distinction between geographic and attribute data, think about your home where you grew up or where you currently live. Within the context of this discussion, we can associate both geographic and attribute data to it. For instance, we can define the location of your home many ways, such as with a street address, the street names of the nearest intersection, the postal code where your home is located, or we could use a global positioning system–enabled device to obtain latitude and longitude coordinates. What is important is geographic data permit us to define the location of an object (i.e., your home) on the surface of the earth.

In addition to the geographic data that define the location of your home are the attribute data that describe the various qualities of your home. Such data include but are not restricted to the number of bedrooms and bathrooms in your home, whether or not your home has central heat, the year when your home was built, the number of occupants, and whether or not there is a swimming pool. These attribute data tell us a lot about your home but relatively little about where it is.

Not only is it useful to recognize and understand how geographic and attribute data differ and complement each other, but it is also of central importance when learning about and using GISs. Because a GIS requires and integrates these two distinct types of data, being able to differentiate between geographic and attribute data is the first step in organizing your GIS. Furthermore, being able to determine which kinds of data you need will ultimately aid in your implementation and use of a GIS. More often than not, and in the age and context of information technology, the data and information discussed thus far is the stuff of computer files, which are the focus of the next section.

Of Files and Formats...

When we collect data about your home, rainforests, or anything, really, we usually need to put them somewhere. Though we may scribble numbers and measures on the back of an envelope or write them down on a pad of paper, if we want to update, share, analyze, or map them in the future, it is often useful to record them in digital form so a computer can read them. Though we won't bother ourselves with the bits

4. Data that describe the qualities and characteristics of a particular phenomena.

and bytes of computing, it is necessary to discuss some basic elements of computing that are both relevant and required when learning and working with a GIS.

One of the most common elements of working with computers and computing itself is the file. Files in a computer can contain any number of things from a complex set of instructions (e.g., a computer program) to a list of numbers and letters (e.g., address book). Furthermore, computer files come in all different sizes and types. One of the clues we can use to distinguish one file from another is the file extension. The file extension refers to the letters that follow the period (".") after the name of the file. Table 3.1 contains some of the most common file extensions and the types of files with which they are associated.

Table 3.1

filename.txt	Simple text file
filename.doc	Microsoft Word document
filename.pdf	Adobe portable document format
filename.jpg	Compressed image file
filename.tif	Tagged image format
filename.html	Hypertext markup language (used to create web pages)
filename.xml	Extensible markup language
filename.zip	Zipped/compressed archive

Some computer programs may be able to read or work with only certain file types, while others are more adept at reading multiple file formats. What you will realize as you begin to work more with information technology, and GISs in particular, is that familiarity with different file types is important. Learning how to convert or export one file type to another is also a very useful and valuable skill to obtain. In this regard, being able to recognize and knowing how to identify different and unfamiliar file types will undoubtedly increase your proficiency with computers and GISs.

Of the numerous file types that exist, one of the most common and widely accessed file is the **simple text**, **plain text,** or just text file. Simple text files can be read widely by word processing programs, spreadsheet and database programs, and web browsers. Often ending with the extension ".txt" (i.e., *filename.txt*), text files contain no special formatting (e.g., **bold**, *italic*, underlining) and contain only alphanumeric characters. In other words, images or complex graphics are not well suited for text

files. Text files, however, are ideal for recording, sharing, and exchanging data because most computers and operating systems can recognize and read simple text files with programs called text editors.

When a text file contains data that are organized or structured in some fashion, it is sometimes called a flat file (but the file extension remains the same, i.e., .txt). Generally, flat files are organized in a tabular format or line by line. In other words, each line or row of the file contains one and only one record. So if we collected height measurements on three people, Tim, Jake, and Harry, the file might look something like this:

Name	Height
Tim	6'1"
Jake	5'9"
Harry	6'2"

Each row corresponds to one and only one record, observation or case. There are two other important elements to know about this file. First, note that the first row does not contain any data; rather, it provides a description of the data contained in each column. When the first row of a file contains such descriptors, it is referred to as a header row or just a **header**. Columns in a flat file are also called fields, **variables,** or **attributes**. "Height" is the attribute, field, or variable that we are interested in, and the observations or cases in our data set are "Tim," "Jake," and "Harry." In short, rows are for records; columns are for fields.

The second unseen but critical element to the file is the spaces in between each column or field. In the example, it appears as though a space separates the "name" column from the "height" column. Upon closer inspection, however, note how the initial values of the "height" column are aligned. If a single space was being used to separate each column, the height column would not be aligned. In this case a tab is being used to separate the columns of each row. The character that is used to separate columns within a flat file is called the delimiter or separator. Though any character can be used as a delimiter, the most common delimiters are the tab, the comma, and a single space. The following are examples of each.

Tab-Delimited	Single-Space-Delimited	Comma-Delimited
Name Height	Name Height	Name, Height
Tim 6.1	Tim 6.1	Tim, 6.1
Jake 5.9	Jake 5.9	Jake, 5.9

Tab-Delimited	Single-Space-Delimited	Comma-Delimited
Harry 6.2	Harry 6.2	Harry, 6.2

Knowing the delimiter to a flat file is important because it enables us to distinguish and separate the columns efficiently and without error. Sometimes such files are referred to by their delimiter, such as a "comma-separated values" file or a "tab-delimited" file.

When recording and working with geographic data, the same general format is applied. Rows are reserved for records, or in the case of geographic data, locations and columns or fields are used for the attributes or variables associated with each location. For example, the following tab-delimited flat file contains data for three places (i.e., countries) and three attributes or characteristics of each country (i.e., population, language, continent) as noted by the header.

Country	Population	Language	Continent
France	65,000,000	French	Europe
Brazil	192,000,000	Portuguese	South America
Australia	22,000,000	English	Australia

Files like those presented here are the building blocks of the various tables, charts, reports, graphs, and other visualizations that we see each and every day online, in print, and on television. They are also key components to the maps and geographic representations created by GISs. Rarely if ever, however, will you work with one and only one file or file type. More often than not, and especially when working with GISs, you will work with multiple files. Such a grouping of multiple files is called a **database**[5]. Since the files within a database may be different sizes, shapes, and even formats, we need to devise some type of system that will allow us to work, update, edit, integrate, share, and display the various data within the database. Such a system is generally referred to as a database management system (DBMS). Databases and DBMSs are so important to GISs that a later chapter is dedicated to them. For now it is enough to remember that file types are like ice cream—they come in all different kinds of flavors. In light of such variety, Section 3.2 "Data about Data" details some of the key issues that need to be considered when acquiring and working with data and information for GISs.

5. A collection of multiple files used to collect, organize, and analyze data.

KEY TAKEAWAYS

- Data refer to specific facts, measurements, or characteristics of objects and phenomena of interest.
- Information refers to knowledge of value that is obtained from the analysis of data.

EXERCISES

1. What is the difference between data and information?
2. What are the differences between spatial and attribute data?
3. Identify each of the files in Table 3.1 according to their extension.
4. Search for and download three different simple text or flat files. Open them in a word processor and spreadsheet program. Use the search and replace function to change the delimiters (e.g., from commas to tabs or vice versa).
5. The US Bureau of Census distributes geospatial data as TIGER files. What are they?
6. Identify resources and websites on the Internet that can help you make sense of file extensions.

3.2 Data about Data

LEARNING OBJECTIVE

1. The objective of this section is to highlight the difference between primary and secondary data sources and to understand the importance of metadata and data standards.

Consider the following comma-delimited file:

city, sun, temp, precip

Los Angeles, 300, 70, 10

London, 50, 55, 40

Singapore, 330, 80, 60

Looking at the contents of the file, we can see that it contains data about the cities of Los Angeles, London, and Singapore. As noted, each field or attribute is separated by a comma, and the file also contains a header row that tells us about the data contained in each column. Or does it? What does the column "sun" refer to? Is it the number of sunny days this year, last year, annually, or when? What about "temp"? Does this refer to the average daytime, evening, or annual temperature? For that matter, how is temperature measured? In Celsius? Fahrenheit? Kelvin? The column "precip" probably refers to precipitation, but again, what are the units or time frame for such measures and data? Finally, where did these data come from? Who collected them, when were they collected and for what purpose?

It is amazing to think that such a small text file can lead to so many questions. Now let's extend the example to a file with one hundred records on ten variables, one thousand records on one hundred variables or better yet, ten thousand records on one thousand variables. Through this rather simple example, a number of general but central issues that are related to data emerge. Such issues range from the relatively mundane naming conventions that are used to identify individual records (i.e., rows) and distinguish one field (i.e., column) from another, to the issue of providing documentation about what data are included in a given file; when the

data were collected; for what purpose are the data to be used; who collected them; and, of course, where did the data come from?

The previous simple text file illustrates how we cannot and should not take data and information for granted. It also highlights two important concepts with regard to the source of data and to the contents of data files. With regard to data sources, data can be put into one of two distinct categories. The first category is called **primary data**[6]. Primary data refer to data that are collected directly or on a firsthand basis. For example, if you wanted to examine the variability of local temperatures in the month of May, and you recorded the temperature at noon every day in May, you would be constructing a primary data set. Conversely, **secondary data**[7] refer to data collected by someone else or some other party. For instance, when we work with census or economic data collected and distributed by the government, we are using secondary data.

Several factors influence the decision behind the construction and use of primary data sets versus secondary data sets. Among the most important factors are the costs associated with data acquisition in terms of money, availability, and time. In fact, the data acquisition and integration phase of most geographic information system (GIS) projects is often the most time consuming. In other words, locating, obtaining, and putting together the data to be used for a GIS project, whether you collect the data yourself or use secondary data, may indeed take up most of your time. Of course, depending on the purpose, availability, and need, it may not be necessary to construct an entirely new data set (i.e., primary data set). In light of the vast amounts of data and information that are publicly available, for example, via the Internet, the cost and time savings of using secondary data often offset any benefits that are associated with primary data collection.

Now that we have a basic understanding of the difference between primary and secondary data, as well as the rationale behind each, how do we go about finding the data and information that we need? As noted earlier, there is an incredibly vast and growing amount of data and information available to us, and performing an online search for "deforestation data" will return hundreds—if not thousands—of results. To overcome this data and information overload we need to turn to...even more data. In particular, we are looking for a special kind of data called **metadata**[8]. Simply defined, metadata are data about data. At one level, a header row in a simple text file like those discussed in the previous section is analogous to metadata. The header row provides data (e.g., names and labels) about the subsequent rows of data.

Header rows themselves, however, may need additional explanation as previously illustrated. Furthermore, when working with or searching through several data

6. Data that are collected firsthand.

7. Data that are collected by someone else or a different party.

8. Data and information that describe data.

sets, it can be quite tedious at best or impossible at worst to open each and every file in order to determine its contents and usability. Enter metadata. Today many files, and in particular secondary data sets, come with a metadata file. These metadata files contain items such as general descriptions about the contents of the file, definitions for the various terms used to identify records (rows) and fields (fields), the range of values for fields, the quality or reliability of the data and measurements, how the data were collected, when the data were collected, and who collected the data. Though not all data are accompanied by metadata, it is easy to see and understand why metadata are important and valuable when searching for secondary data, as well as when constructing primary data that may be shared in the future.

Just as simple files come in all shapes, sizes, and formats, so too do metadata. As the amount and availability of data and information increase each and every day, metadata play a critical role in making sense of it all. The class of metadata that we are most concerned with when working with a GIS is called **geospatial metadata**[9]. As the name suggests, geospatial metadata are data about geographical and spatial data. According to the Federal Geographic Data Committee (FGDC) in the United States (see http://www.fgdc.gov), "Geospatial metadata are used to document geographic digital resources such as GIS files, geospatial databases, and earth imagery. A geospatial metadata record includes core library catalog elements such as Title, Abstract, and Publication Data; geographic elements such as Geographic Extent and Projection Information; and database elements such as Attribute Label Definitions and Attribute Domain Values." The definition of geospatial metadata is about improving transparency when it comes to data, as well as promoting standards. Take a few moments to explore and examine the contents of a geospatial metadata file that conforms to the FGDC here.

Generally, standards refer to widely promoted, accepted, and followed rules and practices. Given the range and variability of data and data sources, identifying a common thread to locate and understand the contents of any given file can be a challenge. Just as the rules of grammar and mathematics provide the foundations for communication and numeric calculations, respectively, metadata provide similar frameworks for working with and sharing data and information from various sources.

The central point behind metadata is that it facilitates data and information sharing. Within the context of large organizations such as governments, data and information sharing can eliminate redundancies and increase efficiencies. Moreover, access to data and information promotes the integration of different data that can improve analyses, inform decisions, and shape policy. The role that metadata—and in particular geospatial metadata—play in the world of GISs is critical and offers enormous benefits in terms of cost and time savings. It is

9. A special class of metadata that contains information about the geographic qualities of a data set.

precisely the sharing, widespread distribution and integration of various geographic and nongeographic data and information, enabled by metadata, that drive some of the most interesting and compelling innovations in GISs and the broader geospatial information technology community. More important, widespread access, distribution, and sharing of geographic data and information have important social costs and benefits and yield better analyses and more informed decisions.

KEY TAKEAWAYS

- Primary data refer to data that are obtained via direct observation or measure, and secondary data refer to data collected by a different party.
- Data acquisition is among the most time-consuming aspects of any GIS project.
- Metadata are data about data and promote data exchange, dissemination, and integration.

EXERCISES

1. What are the costs and benefits of using primary data instead of secondary data?
2. Refer to the Federal Geographic Data Committee website (http://www.fgdc.gov) and describe in detail what information should be included in a metadata file. Why are metadata and standards important?

3.3 Finding Data

LEARNING OBJECTIVE

1. The objective of this section is to identify and evaluate key considerations when searching for data.

Now that we have a basic understanding of data and information, where can we find such data and information? Though an Internet search will certainly come up with myriad sources and types of data, the hunt for relevant and useful data is often a challenging and iterative process. Therefore, prior to hopping online and downloading the first thing that appears from a web search, it is useful to frame our search for data with the following questions and considerations:

1. What *exactly* is the purpose of the data? Given the fact the world is swimming in vast amounts of data, articulating why we need (or why we don't need) a given set of data will streamline the search for useful and relevant data. To this end, the more specific we can be about the purpose of the needed data, the more efficient our search for data will be. For example, if we are interested in understanding and studying economic growth, it is useful to determine both temporal and geographic scales. In other words, for what time periods (e.g., 1850–1900) and intervals (e.g., quarterly, annually) are we interested, and at what level of analysis (e.g., national, regional, state)? Oftentimes, data availability, or more specifically, the lack of relevant data, will force us to change the purpose or scope of our original question. A clear purpose will yield a more efficient search for data and enables us to accept or discard quickly the various data sets that we may come across.
2. The second question we need to ask ourselves is what data already exist and to what data do we have access already? Prior to searching for new data, it is always a good idea to take an inventory of the data that we already have. Such data may be from previous projects or analyses, or from colleagues and classmates, but the key point here is that we can save a lot of time and effort by using data that we already possess. Furthermore, by identifying what we have, we get a better understanding of what we need. For instance, though we may already have census data (i.e., attribute data), we may need updated geographic data that contains the boundaries of US states or counties.

3. Next, we need to assess and evaluate the costs associated with data acquisition. Data acquisition costs go beyond financial costs. Just as important as the financial costs to data are those that involve your time. After all, time is money. The time and energy you spend on collecting, finding, cleaning, and formatting data are time and energy taken away from data analysis. Depending on deadlines, time constraints, and deliverables, it is critical to learn how to manage your time when looking for data.
4. Finally, the format of the data that is needed is of critical importance. Though many programs can read many formats of data, there are some data types that can only be read by some programs and some programs that require particular data formats. Understanding what data formats you can use and those that you cannot will aid in your search for data. For instance, one of the most common forms of geographic information system (GIS) data is called the **shapefile**[10]. Not all GIS programs can read or use shapefiles, but it may be necessary to convert to or from a shapefile or some other format. Hence, as noted earlier, the more data formats with which we are familiar, the better off we will be in our search for data because we will have an understanding of not only what we can use but also what format conversions will need to be made if necessary.

All these questions are of equal importance and being able to answer them will assist in a more efficient and effective search for data. Obviously, there are several other considerations behind the search for data, and in particular GIS data, but those listed here provide an initial pathway to a successful search for data.

As information technology evolves, and as more and more data are collected and distributed, the various forms of data that can be used with a GIS increases. Generally, and as discussed previously, a GIS uses and integrates two types of data: geographic data and attribute data. Sometimes the source of both geographic and attribute data are one in the same. For instance, the US Bureau of Census (http://www.census.gov) distributes geographic boundary files (e.g., census tract level, county level, state level) as well as the associated attribute data (e.g., population, race/ethnicity, income). What's more is that such data are freely available at no charge. In many respects, US census data are exceptional: they are free and comprehensive. If only all data were free and comprehensive!

Obviously, each and every search for data will vary according to purpose, but data from governments tend to have good coverage and provide a point of reference from which other data can be added, compared, and evaluated. Whether you need satellite imagery data from the National Aeronautics and Space Administration (http://www.nasa.gov) or land use data from the United States Geological Survey

10. A common set of files used by many geographic information system (GIS) software programs that contain both spatial and attribute data.

(http://www.usgs.gov), such government sources tend to be reliable, reputable, and consistent. Another key element of most government data is that they are freely accessible to the public. In other words, there is no charge to use or to acquire the data. Data that are free to use are generally called **public data**[11].

Unlike publicly available data, there are numerous sources of **private** or **proprietary data**[12]. The main difference between public and private data is that the former tend to be free, and the latter must be acquired at a cost. Furthermore, there are often restrictions on the redistribution and dissemination of proprietary data sets (i.e., sharing the purchased data is not allowed). Again, depending on the subject matter, proprietary data may be the only option. Another reason for using proprietary data is that the data may be formatted and cleaned according to your needs. The trade-off between financial cost and time saved is one that must be seriously considered and evaluated when working with deadlines.

The search for data, and in particular the data that you need, is often the most time consuming aspect of any GIS-related project. Therefore, it is critical to try to define and clarify your data requirements and needs—from the temporal and geographic scales of data to the formats required—as clearly as possible and as early as possible. Such definition and clarity will pay dividends in your search for the right data, which in turn will yield better analyses and well-informed decisions.

KEY TAKEAWAY

- Prior to searching for data, ask yourself the following questions: Why do I need the data? At what time scale do I need the data? At what geographic scale do I want the data? What data already exist? What format do I need the data?

EXERCISES

1. Identify five possible sources for data on the gross domestic product (GDP) for the countries in Africa.
2. Identify two sources for geographic data (boundary files) for Africa.
3. What kind of geographic data does the United Nations provide?

11. Data that can be shared and distributed freely.

12. Data that must be purchased and are subject to certain terms of use.