5 DATA MANAGEMENT

TIM TRAINOR

5.1 INTRODUCTION

Data comprises the basic ingredients for the content and design of an atlas regardless of the form it takes. Any data or information is considered for this chapter. This includes data that is visual as well as data that may be required to create information as part of the content of an atlas. Examples of types of data that are useful to modern atlases include geospatial, text, images, remotely sensed data, and animation.

Most atlases tend to be at medium and small scale. The granularity of the data will change based on the scale of the atlas. It is important to acquire and use appropriate data for the chosen scale. One basic tenet of cartography is that a map is designed for the final output device. In modern cartography, this tenet is a challenge as the final output device is not always known. Decisions made about the data need to support the design of the maps. This chapter details the differences between spatial data types. Then a discussion of attribute data follows. The chapter closes with a detailed account of the importance of metadata when developing an atlas.

5.2 CARTOGRAPHIC DATABASES

Data is often stored in a way that optimizes collection and update activities in a database.
When using a database for cartographic production it is beneficial to reorganize and classify the data to make it ready for mapping. To make the mapping process easier, cartographic databases are populated potentially with data from a variety of data sources such as other spatial databases. Once in the cartographic database, theses the various data types are likely manipulated from their original state.

The cartographic database predominantly stores features such as transportation (roads, railroads) hydrography (oceans, rivers, lakes), physical features (landforms, elevation) and cultural features (urban areas, buildings, airports). This data is generalized through one of several operations such as aggregating or chaining or dissolving. These processes optimize the data for mapping throughput and occur only once in the cartographic database negating the need for redundant processing with each map made. The data manipulations and database design differences are used to create very diverse and potentially complex maps. When producing a significant number of maps

FIGURE 5-1:

DATA IN A SPATIAL AND A CARTOGRAPHIC DATABASE

on a schedule, performance is paramount and a cartographic database supports higher performance than would be experienced in manipulating data from a common spatial database that supports different purposes. Figure 5-1 shows how data is stored in different ways in a spatial database and a cartographic database. In a spatial database the data are stored as a point, line or polygon. In the cartographic database these points, lines, and polygons have been aggregated, chained, or dissolved to make recognizable features.

5.3 SPATIAL DATA

Spatial data can be classified into two different types of data, vector data and raster data (Figure 5- 2). Vector data is stored as series of x,y coordinates. The vector data model is best suited for displaying discrete data that have definable boundaries. These features are represented in three ways: points, lines, and polygons. Points are defined by a set of x,y coordinates, and are often shown as a marker symbol depicting objects such as the centroid of a building. Points are sometimes referred to as nodes. Lines, also known as arcs, segments, or edges, possess a length equal to the distance between their start and end points. An example of a linear feature would be a road that has a defined beginning and end. Polygons are represented as two-dimensional features and have perimeter and area. Polygons also are known as area or faces. Polygons have the same start and end point and their shape is defined by numerous nodes along the perimeter. The shape of a lake is an example of a polygon feature. The use of vector data is ideal when topological relationships must be maintained (more about topology later). Vector data can readily be linked to attribute or non-spatial data (described below).

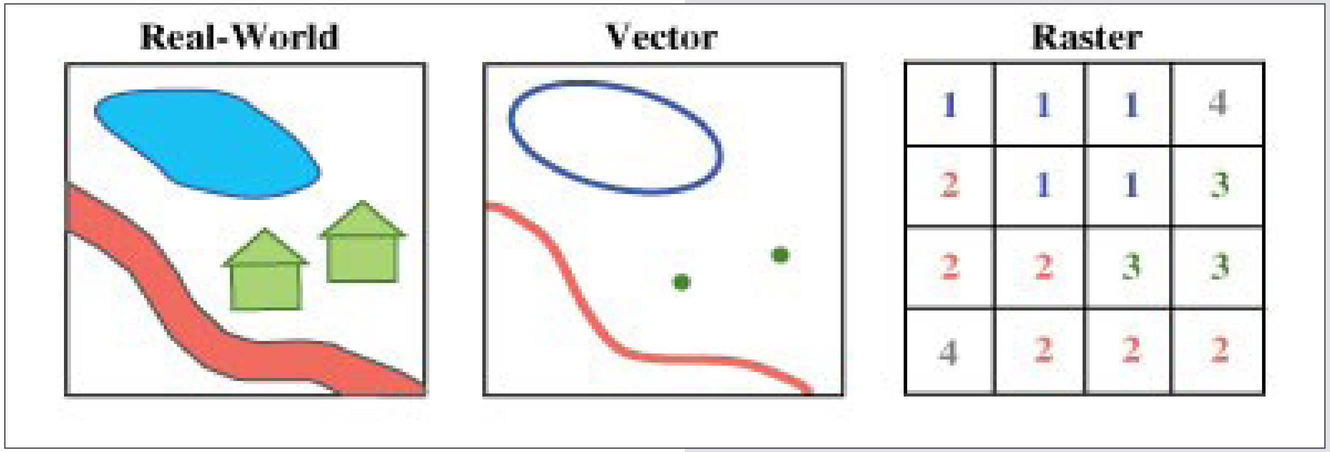
The raster data model uses a grid or matrix to display values. Each cell, or pixel, in the grid is a uniform size and represents one data value that is stored in a table. The size of the grid corresponds to the level of detail desired for the geographical feature. This model is suitable for displaying continuous data that vary without interruption across a geographical space. Values may be measured directly or derived from nearby locations. Temperature and elevation are examples of continuous phenomena. Satellite and areal imagery, and digital elevation models (DEMs) are common examples of raster data used in digital mapping.

5.4 TOPOLOGY

Topology is a data structure based on shared geometry of points, lines, and polygons. There are many benefits to using a topological data structure over a non-topological structure in map creation. Simply, topology simplifies data storage and assures correct relationship among the various data and data types. Topology is maintained through database rules. Topology uses less storage space since boundaries that are shared between adjacent polygons are stored only once. Also, it helps

FIGURE 5-2:

DATA IN THE REAL WORLD, IN VECTOR FORMAT, AND IN RASTER FORMAT



to prevent and correct digitizing errors such as undershooting and overshooting, and provides more effective analysis of adjacency and connectivity, such as allowing for automated map scaling through which scale is determined by point feature proximity. With this automated scaling, points are aware of their relationships and distances from surrounding points (Theobald, 2001).

The advantages of a topological database are numerous, but there are also advantages to non-topological data structures. Chains in a non-topological feature are stored in a sequential manner, whereas with topology this is not a requirement. Therefore, non-topological graphics typically load more quickly. Also, it can be easier to analyze data in a non-topological environment when many attributes exist at the same location. In this situation, there can be many overlapping non-topological layers, each with its own data. Topology is more difficult to manipulate since there is only one layer with multiple attributes for that position. For example, for an unambiguous analysis of different elk ranges in an elk habitat, each range can be displayed as a non-topological, overlapping polygon.

Table 1: Attribute tables for non-topological and topological data structures in Figure 3. Source: Theobald, 2001.

(a) The attribute table for non-topological representation of Figure 5-3. Each polygon has one attribute and overlapping relationships are not indicated.

Polygon	Habitat
1	Winter
2	Severe
3	Concentration

(b) The attribute table for the topological representation of Figure 3. Note that because each polygon can have any of the three attributes (winter, severe, concentration area), each polygon has multiple attributes.

Polygon	Winter	Severe	Concentration
1	yes	no	no
2	yes	yes	no
3	no	no	yes
4	yes	no	no

FIGURE 5-3:

SOURCE: THEOBALD, 2001

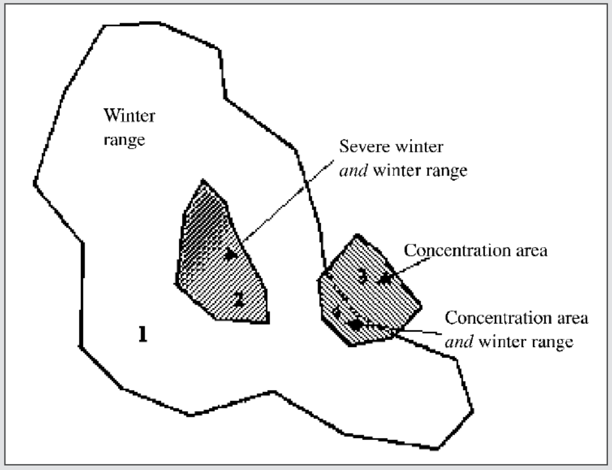
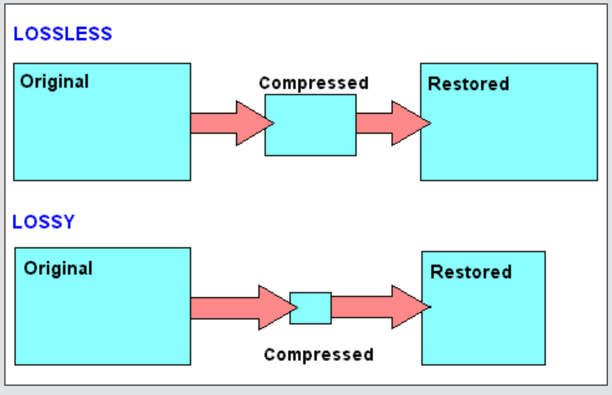


FIGURE 5-4

LOSSY COMPRESSION REDUCES THE QUALITY OF THE IMAGE.



5.5 DATA COMPRESSION

Both raster and vector data are appropriate to use in an atlas, but the amount of computer space that they occupy can vary greatly. Raster images are comprised of many grid cells, with each cell having its own associated value. Because of this format, raster, in its native form, occupies more computer space than vector data. Despite their size, raster data processes more quickly than vector data and may be more beneficial to use when dealing with a large number of datasets that necessitate computational speed (Hohl, 1997).

Data compression is a method for storing raster data more effectively. Raster data compression reduces redundancy in the data, allowing the file to occupy less computer space. The amount of file size reduction depends on the image type, with bi-level images decreasing to approximately 10 percent of the original file size after standard compression, and pseudo color and grayscale images decreasing to about 30 percent of their initial size. When the choice is made to compress the data, it is important to remember that certain types of compression, such as with GIF and JPEG compression, will lessen the quality and appearance of the atlas image (Hohl, 1997).

If attributes were assigned to pixels in a raster image just as attributes are assigned to vector data features, then the raster image would increase by more than twice its original size and data compression would be of no benefit. The file size of vector data is not as sensitive to increases in attributed features because vectors require only one attribute record for each line feature. A raster image requires one record per pixel in the line image. For example, for a vector line feature, it would be simple to define the line as a stream and to assign attributes such as stream type and stream name. This is much more efficient than assigning these same attributes to every pixel that makes up the stream (Hohl, 1997).

5.6 DATA SOURCES

The availability of geographic data varies greatly from one country to another (Crompvoets and Bregt, 2007). Digital geographic data are sometimes provided via storage mediums such as CD or DVD, but, increasingly, dissemination is provided over the Internet. World Wide Web sites sometimes take the form of geoportals, which are access points to geographic data (Tait, 2005). These portals, or clearinghouses, have been developed by government agencies, universities, private companies, and individuals (Goodchild et al., 2007).

Common datasets available are administrative boundaries, land use/land cover, elevation, geologic data, population/demographics, economic data, populated places, roads/transportation networks, hydrography, and imagery. An important element of geoportals is the capability to search for data based on vintage, location, and thematic attributes (Maguire and Longley, 2005). Searches are made possible by metadata which provide information about geographic data, including what the data are about, when and how the data were