5

Data Storage and Representation

5.1 STORAGE MEDIA

Computer cartography demands a familiarity with the tools and limitations of digital technology. In Chapter 4 we examined the conversion of map data to digital form using geocoding. When a map consists solely of numbers, the cartographer is faced with many decisions about how to store the map data and how to select methods of representation that allow the data to be used for mapping. In Chapter 2 we surveyed the various types of devices used in computer cartography, especially graphic input and output devices, but we explicitly left out storage devices. The ultimate purpose of geocoding and choosing a representational method and data structure for cartographic data is to store the digital map on some kind of permanent storage device. Ideally, the data should be easy to get back from storage and should also be stored in a way that is both permanent (usable for more than 10 years, at least!) and transportable to other computers. The means of storage is often also the means of distribution. In the days of paper maps, one could purchase the paper over the counter or through the mail, and the information was contained in the printed inks on the paper's surface. Digital map data have to be distributed also. In this chapter we deal with the means of storage, both physical and logical. In Chapter 6 we examine more closely the role of distribution of digital map information.

Over time, improvements in storage technology have drastically reduced the cost and space requirements for storage, and permanence has improved. A key issue for storage remains portability between computers, although distributed computing and networks have relegated this problem to one of archiving. During the initial years of computing, the most widely used storage methods were the punched card and paper tape. These storage methods were not permanent, and they suffered from additional problems, such as sequencing and tearing. Disk and magnetic tape storage replaced cards and paper tape. Fixed or hard disks are usually part of the computer and as such are not transportable, unlike their smaller but portable equivalent, the removable disk. Removable disks, however, are fragile and bulky and can be damaged by strong magnetic fields.

Floppy disks, introduced when microcomputers became available, are relatively rigid, so a preferable term in use is *diskette*. Older disks were enclosed in a plastic envelope containing the thin magnetic sheet that holds the data. Early types were 133 mm (5.25 inches) and 203 mm (8.0 inches) in diameter, but they are rapidly being replaced by 89-mm (3.5-inch) disks, which have a more rigid plastic case and can support higher storage densities. Such a disk has been used to distribute the software and data for use with this book, and is enclosed in a plastic pocket at the back of the book.

Probably the longest surviving and most reliable storage medium is magnetic tape, which comes both as loose reels and tape cartridges. Different lengths of tape and different numbers of bits per inch (BPI) of storage density allow different amounts of data to be stored. The cost, reliability, transferability, and high density of magnetic tape made this the normal medium for data and software distribution for large computers. The advent of workstations has led to several new tape formats, including the 4-mm tape cartridge and the 8-mm tape. The latter are highly flexible and extremely compact and are capable of storing between 2.3 and 5.0 gigabytes of data, depending on blocking and structure, at a cost of only a few dollars. This change clearly indicates that basic level digital cartography now faces few physical limitations of storage volume, even on microcomputers.

More recently, mass storage has become possible using optical disk technology. Many optical disks are read-only and can play all or part of a record but cannot record. The first generation of optical disks used compact discs as read-only memory, and disk templates had to be written in much the same way that a book is printed. More recent optical media allow both reading and writing. The advantage of optical media is the size of the storage, sometimes approaching gigabytes of storage even for a microcomputer. Almost all topographic map coverage for a single state, both current and historical, would fit onto a single optical disk. Several companies now distribute atlases, gazetteers, and even street maps in CD-ROM, such as the Map Explorer software and data by DeLorme. The disks themselves are comparatively fragile and require special holders for placement into a drive. Nevertheless, the prices are now so low that they are a favored means of data and software distribution. The per item costs for disks that are sold in volume can approach the cost of cartridge tape. For example, many government agencies distribute CD-ROM data sets for about $30.

Recently, optical disk “juke-boxes” have become available. These devices allow the user to select, through software, which CD-ROM is to be loaded onto the drive to be read. Stacks of CD-ROMs are therefore available to the user, making hundreds of gigabytes accessible. So large have storage volumes now become that discussions use the terms *terabyte* (1,000 gigabytes) and even *petabyte* (1,000,000 gigabyte). The archive of Landsat data at the EROS Data Center in Sioux Falls, South Dakota, for example, now contains about 40 terabytes or 0.04 petabytes. Perhaps a petabyte of storage for a desktop cartographic workstation may be possible, perhaps sooner than we think! Such a situation ensures a vast redundancy in the supply of digital map data, guaranteeing its survival in the same way as publishing ensures the survival of the information content of a book. Similarly, every cartographer may eventually have access to all the maps and images ever created.

Even reliable storage such as disk memory has to be backed up to a more permanent storage medium to guard against failure and loss. Backups are either full-system or incremental, in which every new or changed file on the entire system is copied. All files are copied onto magnetic tape, and the tapes are archived. If anybody inadvertently deletes a file, or if something goes wrong with the computer, data can be retrieved as they were saved on a given day. Magnetic tape readers are available for microcomputers and are usually used to back up a hard disk on magnetic tape. Even microcomputer hard disks must occasionally be backed up to diskette or tape, which are more likely to survive catastrophic events than hard disks.

5.2 INTERNAL REPRESENTATION

The various storage media in use are simply the physical mechanisms for storing digital cartographic (and other) data. To gain insight into computer cartography, we still have to understand how it is that the data are translated into the physical storage properties of the particular medium, be it a magnetic tape, an optical disk, or a diskette. A basic problem is that although a programmer or a cartographer may seek to access data at random, or by a spatial property, nevertheless on the physical storage device the data are stretched out sequentially, one after the other. Using any storage medium involves keeping track of information about where different data are stored on that disk, or tape, or whatever the medium used. This involves blocking, or dividing the data into manageable chunks, and requires a mechanism for dividing the disk into blocks and sectors. This blocking is independent of any spatial data structure that we may impose upon the digital cartographic data, and even of the file structure used to structure data.

A diskette has radial sectors broken up into blocks or areas reserved for storage of different kinds. The operating system is designed to keep track of this information for you. The link between the actual "map" of the blocks, the sectors, and the physical disk itself is the directory. The *directory* is an area of the disk reserved for information about where other things are on the disk. All operating systems have a command to allow you to see what files occupy your storage, although most mask from the user the specifics about where the storage is actually located. Many computer operating systems allow you to have access to much more storage of one kind, such as RAM, than is apparently available by connecting memory of another type, such as disk for RAM. This method is called virtual memory.

Addresses are locations in storage, both permanent and temporary, where information can be stored. The larger the computer's memory, the larger the range of addresses available for storage of information. The very lowest level of address normally contains the operating system. The basic storage unit is the lowest-level piece of information we can fit into or retrieve from an address. On an optical storage system, addresses relate to resolution. On a mechanical or electrical system, addresses are related to the ability to hold a magnetic or electrical charge. Storage devices are state storage mechanisms; they can store on or off, binary 1 or 0, or one bit. A capacitor can hold a charge or not. An optical reader may see a bar code with a black line or a white line.

The mathematics that corresponds to this bit logic is called *Boolean mathematics,* and it works in number base two, the counting mechanism we would use if we had only one

finger on each hand. Counting is in the sequence of zero, one. As we string the bits together we can build bigger and bigger numbers. There is a clear correspondence between distribution of digits in binary numbers and what the memory storage actually looks like. On the older storage media, such as cards and paper tape, you could actually see the binary representations as rows of holes punched out of the paper. On magnetic media, the tape either holds a charge or does not.

As an example, the binary number 1010 0011 0001 corresponds to the decimal number 2,609. Notice that just as we often break off the 2 from the 609 with a comma for ease of interpretation (most other nations use a space, an apostrophe, or a period), we use spaces between groups of four binary digits or bits. This has the distinct advantage that it is easy to represent the number in other number bases. The two most common alternative number bases are base eight (octal) and base sixteen (hexadecimal). This is because one octal digit corresponds directly to three bits, while one hexadecimal digit corresponds directly to four bits.

By far the most used is hexadecimal, in which the counting numbers are 0, 1, 2, 3, 4, 5, 6, 7, 8 ,9 , A, B, C, D, E, and F. Each cluster of four bits can be represented as a single hexadecimal digit. Thus 1010 translates to decimal 10 or hexadecimal A; 0011 translates to decimal 3 or hexadecimal 3; and 0001 translates to decimal 1 and hexadecimal 1 (Figure 5.1). So the binary number above can be represented as the three hexadecimal digits A31.

To write the binary equivalent of this value onto the storage medium, many different systems are used. Some write the bit values left to right, some right to left, and some invert the value of the bits byte by byte, a process known as byte swapping. The differences relate to whether the least significant bit (that is, right-most logically) is stored first or last. A cluster of four bits can be represented by a single hexadecimal digit. Two hexadecimal digits together are called a byte. We use the term *word* to refer to the grouping of bytes which the computer itself uses as the basis for storage and computation. Thus we

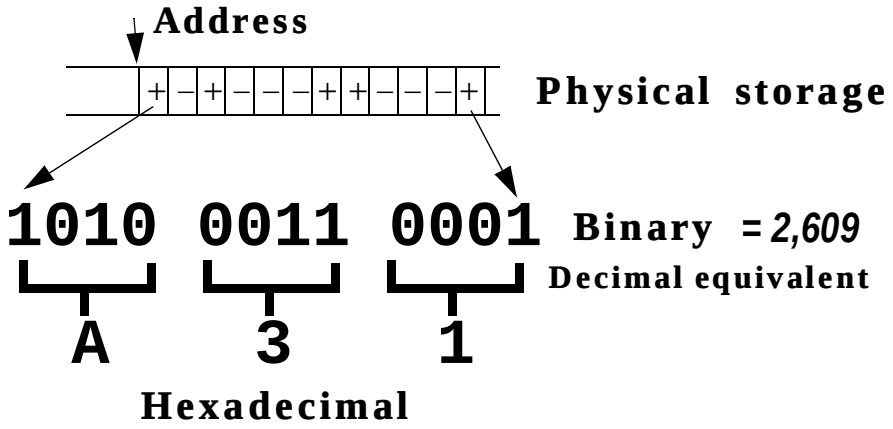

Figure 5.1 Decimal, binary and hexadecimal representation of the number 2,609.

may hear that certain chips are 16-bit or that a particular computer has a 32-bit word. In the early days of microcomputers, there was a big difference in the size of a word on a microcomputer and the size of a word on a large computer. There were even differences company by company on what the size of a word was on a large computer. Today, even some small computers use 32-bit words, while the very largest and most powerful computers have moved on to 64-bit words or other sizes. A larger word size allows a greater range of addressing and more precision on computations.

Because hexadecimal is an intermediate internal representation of the data and programs that are stored on a computer, many computer users never need know this level of detail. Similarly, we may be able to drive a car without opening the hood. When the details under the hood become important is when something is wrong with the car, when we decide to build a new car or to improve the one we have. Programmers make frequent use of binary and hexadecimal, but as a computer user the only time you may see hexadecimal is when something goes wrong. When a program or operating system crashes, it sometimes tries to send a message to assist in finding what went wrong. What the computer sends in the message is an address, or a location in memory, where the error occurred, and usually it gives the address in hexadecimal.

Some scientific pocket calculators do binary, octal, and hexadecimal arithmetic. Using these calculators is a good way to get some experience in using these number systems, although using them for shopping or balancing a checkbook can be confusing. Number operations combine binary digits with operators called AND, OR, and NOT and their inverse combinations (NAND and NOR). AND simply matches two binary numbers and results in a 1 only if both bits are 1. OR results in a 1 if either bit is 1. NOT simply results in the opposite of a bit, zero replaces one and vice versa. Combinations of these operations between numbers stored in memory locations—called registers—are how the computer performs arithmetic.

Hexadecimal codes can be used to number the available storage locations in memory, which are inherently sequential, so that they can be matched with their data contents. Having found the storage location, however, we may wish to retrieve what we find there or change the memory to what we want. If the memory locations themselves held hexadecimal digits, then the computer would only be able to store numbers, and only positive numbers at that.

To get over this limitation, we have developed a standard way of encoding the more complex sequences that we actually use, numbers, characters (such as upper- and lowercase letters), and special characters (such as brackets and exclamation points). We do not actually store the letter *e*, or the number *0*, we store a representation of them in these basic storage codes. What gets stored in the memory locations pointed to by hexadecimal addresses are binary strings that correspond to character sets.

Generally, the most commonly accepted character set is the ASCII (American Standard Code for Information Interchange) set. The ASCII character set stores one character to one byte, giving us 255 possible codes, although only 127 are normally used (that is, 7 of the 8 bits). The first code is called null, or nothing (*not* zero). The next codes correspond to `control a` through `z`. Remember that on a computer the `control` key is just like "shift" on a typewriter, the difference between uppercase a (`A`) and lowercase a (`a`) is to hold down the shift and hit `a`, or do not hold down the shift and hit `a`. `Control`

works in exactly the same way as `shift`. Just as we have upper- and lowercase characters, we have control characters in the code. ASCII control codes go from hexadecimal 01, which is `control-a`, through hexadecimal 1A, which is `control-z`. ASCII code 1B corresponds to the `escape` key. Many escape sequences do special things, such as graphics. When they get ASCII code 1B hexadecimal (27 decimal), many graphics devices interpret the next piece of data as a command rather than data. This explains why listing a binary file on a monitor often produces unexpected results!

ASCII code hexadecimal 20 (decimal 32) is a space; then we go through all the special characters, exclamation points, pound signs, periods, and so forth. At ASCII code hexadecimal 30 (decimal 48) we start with numeral zero and go up through ASCII code hexadecimal 39 (decimal 57), which is numeral nine. At hexadecimal 41 (decimal 65) we start with uppercase a, up to hexadecimal 5A (decimal 90), which is uppercase z. At ASCII hexadecimal 61 (decimal 97) we start with lowercase a, and at hexadecimal 7A (decimal 122), we are at lowercase z.

Internally, therefore, what actually get stored on the storage medium used for computer cartography are the binary representations of ASCII codes. These values are written into hexadecimally referenced locations in either temporary or permanent memory. The advantages of storing ASCII codes are many and include being able to view and edit the data, as well as maintain a high degree of portability between computers. As we will see, however, using ASCII codes often leads to memory management problems, and these codes are not optimal for storing the raw numbers into which digital maps have been converted by geocoding.

5.3 STORAGE EFFICIENCY AND DATA COMPRESSION

An important goal for the storage of data within computers is to achieve storage efficiency. We often have to make a trade-off between storage requirements and the ease of use of cartographic data. Ease of use is reflected by several competing goals: achieving rapid access to data, sustaining fidelity, allowing concurrent access to multiple users, and facilitating data update and maintenance. In Chapter 4 we noted that one of the fundamental characteristics of geographic and cartographic data is sheer volume.

Geographic data sets are typically very large. This means that storage efficiency problems are pertinent to handling cartographic data. Frequently, we are concerned with how we can reconfigure the data to fit into our particular storage device or to fit our logic or reasoning system. To be efficient with storage, we seek to retain as much *cartographic information* with as little storage as possible.

Storage efficiency concerns two sets of storage demands. First, the entire digital cartographic data set must reside in permanent storage at least somewhere on a network and be accessible to the mapping program as required. Second, a computer cartographic program must bring part of or all the map data into the random access memory (RAM) of the computer for display, analysis, and editing. The computer program moves the data from storage into a data structure in RAM supported within the program itself. Data in RAM are available much more rapidly to the program. Data in secondary storage must be brought into and out of the RAM, a process known as I/O, or input and output. To move data between these two types of storage is time consuming. Because digital cartographic

databases are typically large, I/O is the most time-consuming part of producing maps by computer. When only part of the map can fit into the RAM at any given time, such as on a microcomputer that can only access RAM in 640 kilobyte partitions, memory management can become a significant part of the computer mapping software.

The storage efficiency of cartographic data in RAM is determined by the precision of the data and the suitability of the programming data structure. The file size and byte-by-byte character of the data are determined by the physical data structure, (that is, how the logical spatial constructs map onto the computer's memory). The physical structure is the "translator" of the logical structure, the theoretical basis of how the data structure encodes space and spatial relationships, into digits.

A programming data structure is made up of the physical structure into which the data have been put, and the power of the logical data structure as a construct for solving the particular mapping purpose. Obviously, when an effective logical structure is matched by an efficient physical structure, both the analytical and the computer cartographer benefit. The distinction between logical and physical data structures is important and allows different degrees of storage efficiency to be attained.

In Chapter 4 we considered the merits of various graphic input devices. We noted that there are two major types, semiautomatic and automatic, and that historically these structures have influenced programming data structures and storage. In general, the semi-automatic devices produce vector data and the automatic produce raster data. Vector data can be very storage efficient, because we only need to capture information when there is information worth recording.

For example, on the right map projection we can represent the boundary of the state of Colorado with only four points. Curved lines are sometimes a little less efficient because to maximize the amount of information stored we have to concentrate data in areas where the boundary changes in direction. In digitizing a line we capture more points in areas of extreme curvature than in areas that are very straight. This introduces the concept of *information content.* What is a high information content point and what is a low information content point?

Figure 5.2 shows some lines with different degrees of information content at different points. On a straight line, the two end points are very high information points, because they contain information about where the line begins and ends, without which it is impossible to draw the line. A point along the straight line is redundant, because it adds almost nothing to the information content. It does, however, add to the data volume. To add a third x and y value along the straight segment, we have to increase the x and y data storage volume by a third, but we get no corresponding increase in the information content. To achieve storage efficiency, we seek to minimize the redundancy within the data and to maximize the information content. In a vector system we can vary the sampling density to correspond to what we are mapping.

This concept does not apply only to lines, but to points and areas, too. For example, on a surficial geology map, for which we know that there are two basic geology types with one twice as abundant as the other, we would anticipate two different categories for the map area, say schist and gneiss. We know that there is a boundary that runs somewhere within the map area. How would we sample that area to achieve the maximum amount of information content in the data versus the minimum amount of redundancy?

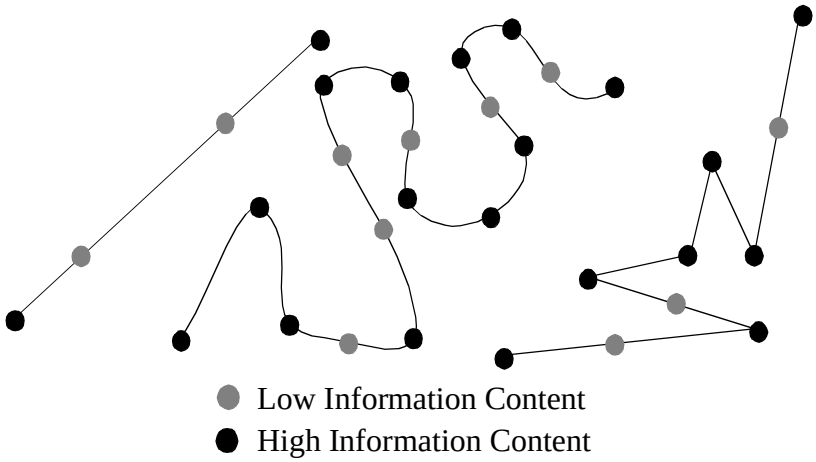

Figure 5.2 Lines with points of different information content

It would be pointless (or rather point-full!) to put a grid over the map and take soil samples at the intersection of every grid point. Most of our data would be redundant , and the sampling strategy would probably miss the only significant feature, the boundary. If we know there is a boundary, we need only a few samples over the map area but many along the boundary itself. This would locate the boundary on the map. Any other sampling strategy would contain redundancy. A vector-based system, in this case irregularly distributed points rather than lines, allows us then to seek out the highest information content areas on the map and to concentrate our sampling effort on those. This strategy, however, assumes some prior knowledge about the nature of the geographic distribution involved.

Raster systems work differently. Using a grid, there is one data value for every grid cell, or grid intersection point per grid. This means that we have a large amount of embedded redundancy and high data volumes for the same geographic area. This is the same as uniform sampling. Within a lake, for example, all grid points simply repeat that this is still a lake.

This strategy is most suited to geographic phenomena that change continuously, such as elevation and physical properties, rather than discontinuous properties, such as surface geology and human characteristics such as rural land-use. Even so, a grid resolution suitable for one geographic region or for a particular phenomenon with distinct geographic properties is usually unsuitable for another. Thus we would use a different grid spacing for elevations on the Great Plains than for elevations on the Rocky Mountains.

Grids happen, however, to be a convenient measurement mechanism. Thus grids are often used for large areas at small scales, while vectors are frequently used for small areas at large scales. The two data structures are really answering two different questions and making two different demands on the data. The two systems are both space-sampling strategies, one regular and the other irregular; one with constant resolution and one with

variable resolution. If we accept a direct relationship between scale and resolution, then vector and raster conversions are a problem of scale transformation. Tobler has suggested the single term *resel*, or resolution element, to apply to both regular and irregular areal sampling distributions, with the *pixel*, a square or rectangular resel, as a special case.

5.3.1 Storing Coordinates

A distinct link between logical and physical data storage is important for storage efficiency. In this chapter we consider the physical aspects of storage. In Chapter 6, we consider the major logical map data structures. We saw in Chapter 4 that under the UTM coordinate system we could represent the location of almost any point on earth to a precision of 1 meter by using a string such as

4,513,410 m N; 587,310 m E; Zone 18, N

During the digitizing stage of geocoding, at some stage we moved a cursor to this point and pressed a button. What geocode does this process transfer from the map to storage? The full reference consists of 15 digits, plus a binary value for the hemisphere (north or south). The whole string, as shown above, is actually 32 characters, 15 digits, four commas, one blank, two semicolons, nine letters, and an end-of-line character to separate this from the next line. Simply storing that the values consist of a UTM reference for zone 18 N as a northing, followed by an easting, reduces the storage to 15 characters

4513410 587310

So by storing a line such as

UTM Zone 18 North, Northing then Easting

only once, for 41 bytes (one ASCII code per character, plus the new-line), we can save 20 bytes per record. For 100,000 records, we save 2 megabytes at a cost of 41 bytes. Let's see if we can do better than that. Zone numbers go from 1 to 60.

When we are designing storage systems, we start by looking at the maximum value we need to fit in any field. We only need to store zone numbers between 1 and 60. For the hemisphere we can store either 1 or 0, one binary digit, instead of a whole byte. What is the maximum easting? If we reached 1,000,000 we would be in the overlap between zones, so for all practical purposes 999,999 is the maximum.

The maximum northing is 10 million. Because 10,000,000 south can be represented as 0 north, and because we cut off at 84 degrees north, the maximum northing is 9,999,999. Thus in reality each record can consist of two numbers, one between 0 and 9,999,999 and one between 0 and 999,999. Using hexadecimal, so that digits correspond to half-bytes, these numbers are 98 96 7F and 0F 42 3F, requiring a total of only six bytes per coordinate pair. Furthermore, if these bytes are written sequentially, that is, without the new-line characters, our 100,000 records need occupy only 600 kilobytes. The data

set, which formerly would have occupied a significant portion of a hard disk, would now fit easily into RAM on a large microcomputer.

If we took the additional step of dropping the last digit, i.e., giving our data a precision of 10 meters and storing the first 100-km-grid-square letter and number designation as shown in Chapter 4, we need store only eight digits, four each for *x* and *y*, making only 10 ASCII characters (bytes) per record, or two values with a maximum of hexadecimal 27 0F, making four bytes per record.

The 100,000 records now take up 1 megabyte as ASCII codes or 400 kilobytes as binary sequential. By simply economizing on the storage representation of a record we have reduced a data file from 3.5 megabytes to 400 kilobytes.

This is an example of achieving storage efficiency by physical data compression, reducing the actual number of bytes required to store information. The storage improvements achieved here, a compression to 11.4% of the (rather padded) original size, were achieved in two ways. First, we used physical compression, achieving savings mostly by using binary instead of ASCII codes. This compression is directly reversible.

Second, we dropped the last digit of the data, saving space but making the compression a noninvertible transformation of the data, actually a scale transformation since we have spatially generalized the data. Dropping the last digit may also create duplicate data points, which can be reduced to a single instance. It is impossible to get this detail back accurately, so a scale change is partly a physical but also partly a logical compression of data storage volume.

The example above is based on reality. The U.S. Geological Survey's land-use and land-cover (LULC) data use a similar compression method. The LULC data in a format known as GIRAS (geographic information retrieval and analysis system) are the digital equivalents of the 1:250,000 land-use and land-cover maps, available for the whole country. In these data sets, which are also topologically encoded, eastings and northings are relative to the nearest 100,000 meter UTM intersection to the west and south of the map area.

This is equivalent to storing once the letter and grid number designations from the Military Grid, covered in Chapter 4. These are stored separately, as headers or "meta-data" (data about data) once per file. The eastings and northings are then only five digits long. In addition, the minimum resolution is set to 10 meters rather than 1 meter; in other words, they lop off the last digit. In storing an arbitrary origin once, the zone number and the hemisphere are stored also. This leaves eight bytes for each point because the files are ASCII. No new-line characters are necessary because the points file is simply written sequentially or in some data structure.

An effective means of achieving optimum storage is to limit the geographic area of coverage. Examples are counties, cities, municipalities, and states. Even within these areas, data use can be optimized by storing data to reflect use, such as by making the most frequently used data the easiest to retrieve from storage.

This is common when data bases are tiled, or divided into squares or rectangles across the area of interest. Another way is to make the data structured by scale and purpose, so that where low-resolution data are suitable for mapping needs, they are used rather than the entire data set.

5.3.2 Compressing Coordinate Storage

Some digital cartographic databases or mapping systems have differential accuracy levels embedded in them. In these systems, for each *x* and *y* value there is another value, an identifier, that says "Include this point at this level." To make a very highly generalized map, we would scan this last character and include the point only if it meets the lowest generalization level, such as 9, which would be the least detailed. A level 9 map would simply exclude all points with level less than 9. A level 1 map would include every single point. Using this method, we can embed different levels of resolution within the same data set, at the cost of storing one more character per point. An effective way to determine the appropriate resolution factor is to apply a line generalization method, such as the Douglas-Peucker technique, discussed in Chapter 10. Generalizing maps can significantly reduce the amount of storage required for them.

Many data compression methods are suitable for all data files, whether spatial or not, whereas others are unique to spatial data and even a particular data structure such as a grid. Many data compression schemes get much of their saving by converting from ASCII into alternative representations. A popular method of physical data compression is Huffman coding. Huffman coding puts an entire file through a one-step process, the result of which can be a significant saving in storage. The process is entirely reversible, and in doing so the reconstructed file is exactly the same as the original. Files with much redundancy in them, many blank lines and columns for example, or large numbers of similar digits compact significantly with Huffman coding. This is often the case with map coordinate files. Huffman coding works on ASCII files and performs a frequency analysis of the ASCII codes. The ASCII codes are then replaced by a new set of Huffman codes of variable length, with the very shortest length codes corresponding to the most frequently occurring ASCII codes (Held, 1983).

Huffman and other compression schemes are often made available to users as part of the operating system or as utilities. MS-DOS 6.0, for example, implements compression of files to approximately double the apparent size of a hard disk. File transfer by diskette often takes advantage of file compression using public domain utilities such as ARC and PKZIP, which can link files together and compress them, allowing a limited-size disk to transfer large files. The Unix operating system contains the utility `compress`, which creates a file with a `.Z` extension of considerably reduced size. Compressed files are particularly useful for transfer of files over networks, where the highest information content per byte during the transfer is desirable. Similarly, the GIF and JPEG file formats use compression schemes for storing images (Kay and Levine, 1992).

There is necessarily a penalty for achieving maximum storage efficiency, and what we usually sacrifice is ease of retrieval, or analytical flexibility. Especially now that the cost and limitations of data storage are falling, the savings of space may not be worth the loss of flexibility or precision used to achieve storage efficiency. As a rule, ASCII data representation is superior for cartographic data unless the mapping is in a production environment, because errors can be detected simply by looking at the data with an editor, even for very large data sets and very rapid browsing.

The human eye and brain are powerful error detectors even for large numbers of records. The most effective error detection (and correction) strategy is when the eye/brain

method is coupled with automatic error-detection routines. Although data compression may be desirable for data transfer through networks, or archiving on tapes, cartographic data are best represented as the actual location coordinates they ultimately must consist of.

Given the basic means by which data can be stored and compressed on computer storage media, which methods are used in computer cartography, and why? In Section 5.4 we examine some of the formats that have been developed and used, especially by government agencies, and determine the significance of the spatial data transfer standard.

5.4 DATA STORAGE FORMATS FOR CARTOGRAPHY

The beauty, and also the problem, of standards is that there are so many to choose from! Because most digital cartographic data archives were developed piecemeal over a long period and for an enormous number of applications, there are a great many proposed standard formats. The need for standardization is most apparent when data must be transported between applications and between computer systems. The most successful data formats are those that have withstood transportation between systems and those that have had the most use and documentation. In this section, we examine in detail the formats used in the past by the major producers of digital cartographic data within the U.S. federal government—the Defense Mapping Agency (DMA), the National Ocean Service (NOS), the United States Geological Survey (USGS)—plus a widely distributed data format, the World Data Bank. The logic of the Bureau of the Census's formats was discussed in Section 4.7. Although data formats related to specific file and byte structures have often become *de facto* standards, they are rarely standard in the more formal sense. In Chapter 6 we examine the spatial data transfer and other formal data standards and how they relate to the future provision of digital map data.

5.4.1 Formats at the U.S. Geological Survey

Digital cartographic data from the U.S. Geological Survey are distributed by the Earth Science Information Centers as part of the National Mapping Program. USGS digital data fall into four categories: *digital line graphs* (DLGs), *digital elevation models* (DEMs), *land-use and land-cover digital data* (GIRAS), and *digital cartographic text* (Geographic Names Information System, GNIS). The USGS continues to improve coverage of the United States and distributes the map data products on computer tape, floppy disk, and CD-ROM.

The DLG data are digital equivalents of the 7.5- and 15-minute USGS sheet maps, as well as the more generalized 1:100,000 maps and the National Atlas regional-level 1:2,000,000 maps. The DEM data are land surface elevations, at both 1:24,000 with a ground spacing of 30 meters using the UTM grid, and 1:250,000 with a ground spacing of 3-arc seconds, the same as the DTED data described later. The land-use and land-cover data are digital versions of the 1:100,000, and 1:250,000 land-use and land-cover and associated maps (the interim land-cover mapping program for Alaska). Finally, the GNIS is a unique data set that includes the text with its attributes from a large number of USGS

map products. By 1994, about 15% of the United States was completely digitized from the 1:24,000 maps, 50% of the 7.5 minute DEMS were finished, and the 1:100,000, GIRAS, and 3-arc second DEMs were entirely finished.

As their name suggests, the digital line graphs are vector encoded and divided by map source scale as separate data sets; the scales are 1:24,000, 1:100,000 and 1:2,000,000. At the 1:24,000 scale, the data consist of boundaries (political and administrative, such as the municipalities, federal lands, and national parks), hydrography (standing water, flowing water, and wetlands), the boundaries of the public land survey system (down to sections in the township and range system), transportation (including roads, trails, railroads, pipelines, and transmission lines), and other significant fabricated structures, such as built-up areas and shopping centers. These maps were digitized from 7.5- or 15-minute quadrangles, or from their compilation products. The 1:100,000 data sets are similarly structured and are subdivided by groups of files covering 30- by 30-minute blocks, from the 1:100,000 scale topographic maps. The 1:2,000,000 maps, digitized from the 1970 National Atlas of the United States of America, follow the same DLG format. Twenty-one digital map sheets cover the United States, fifteen for the coterminous states, five for Alaska, and one for Hawaii. Coverage for New York, for example, includes all the northeastern states, from New York to Maine.

Although these data were originally to consist of three levels of accuracy and coding, in fact all data are provided to the maximum level, that of DLG-3. DLG-3 data are topologically structured and consist of nodes, lines, and areas structured in a manner that explicitly expresses logical geometric relationships. The coded geographic properties are adjacency and connectivity. This allows both plotting of the data for computer cartography and the use of derived information in analytical cartography. The coordinate system is local to the map in thousandths of an inch, but parameters for conversion to UTM are provided in the header files. The latter fact shows clearly how these data relate closely to their sources as paper map products.

The basic cartographic objects in the files are nodes and lines, with areas consisting of labels and links to the other objects. Lines begin and end at a node and are geocoded with the left- and right-hand areas they divide. Lines connect at nodes, and no line crosses itself or another line. Islands are lines that connect at a dummy node; they are termed degenerate lines. Areas are delimited by lines and contain a point, to which is assigned a label for the area. In addition, lines have attribute codes that determine what cartographic entity is being represented. These codes are based on the USGS 7.5-minute map series symbol list. Figure 5.3 shows a typical segment from a DLG and gives examples of different attributes.

An ambitious attempt is now under way to revise the DLG format substantially to encode the spatial relationships between the geographic features on the ground more fully, rather than between their cartographic representations on the map. The new data format, known as DLG-E (Enhanced) will eventually replace the older DLG format. The DLG-E format closely reflects the terminology and data model of the Spatial Data Transfer Standards.

The model is based on features, and includes a set of feature codes for objects such as bridges and rivers as well as a set of topological relationships. As such, DLG-E is more of a move toward the DMAs approach of using integrated topology in map encoding.

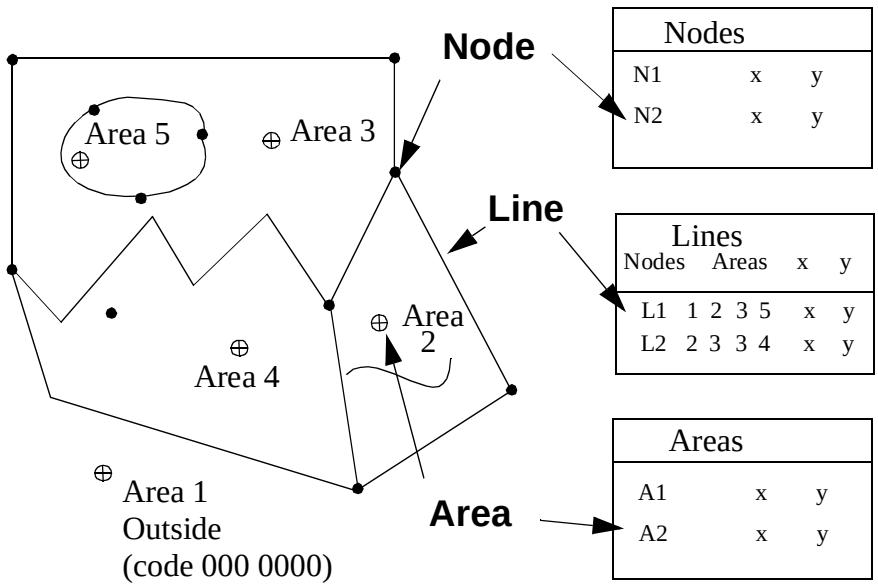

Figure 5.3 Sample digital line graph coding format

Most interesting is the inclusion of text, such as place names, as attributes of features, a significant and important diversion from the current format, in which text is encoded in a separate and unrelated file called GNIS (Geographic Names Information System). A complete discussion of the DLG-E format is contained in Guptill (1990).

The Digital Elevation Models are digitally encoded sample measurements of the surface elevation of topography. Two source map scales are involved, 1:24,000 and 1:250,000, and there are four different map generation technologies. The 1:250,000 data are 3-arc second increments of latitude and longitude for 1-degree blocks covering most of the United States. These data are very similar to the DMA's DTED data and in fact are derived from them. Each 1-degree block contains 1,201 elevation profiles, with 1,201 samples in each profile. Elevations are in meters relative to mean sea level, and latitude/longitude uses the 1972 World Geodetic System datum. Figure 5.4 shows how this data set is logically organized. These sets are distributed on computer tape, with documented headers and formats.

The 1:24,000 data correspond to the 7.5-minute quadrangles and are in UTM coordinates. As such, the boundaries of the map are not square, and the number of elevation samples varies by south-to-north profile over the map. Each profile has a local datum, given in a profile header at the start of each string of samples. The files are ASCII records, stored sequentially on magnetic tape. Figure 5.5 shows this logical format, and it is important to note that the physical structure depends somewhat on which side of the central meridian for the UTM zone the data set falls.

The last USGS data set to be considered here is the digital land-use and land-cover data. Although now largely out of date, these data sets are available by 1-degree blocks

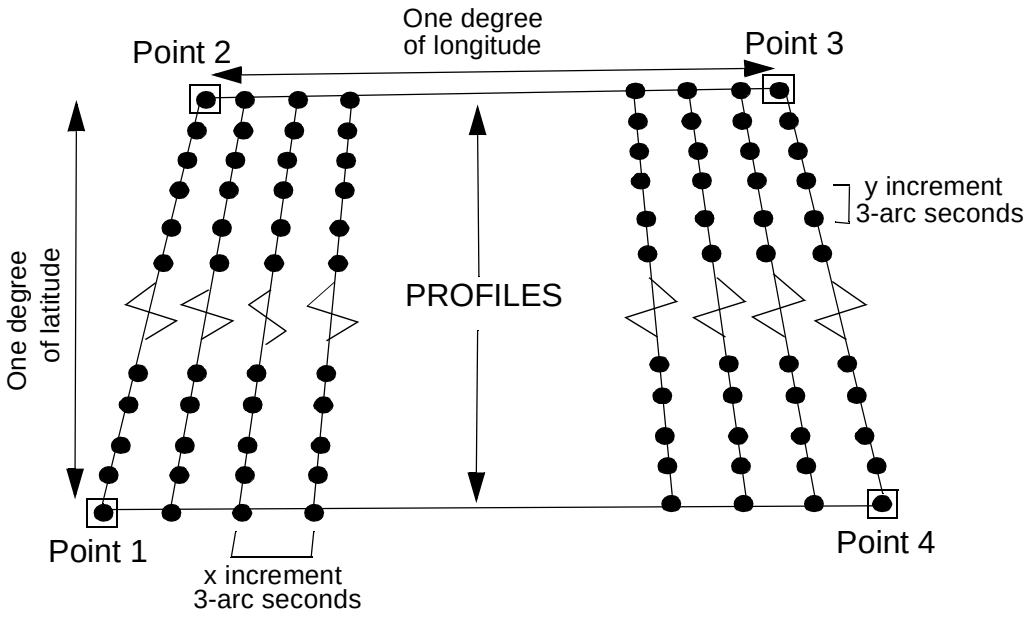

Figure 5.4 1:250,000 3-arc second DEM format from the USGS.

for the entire country and were digitized from the land-use and land-cover sheet maps, with source dates going back to the late 1970s. The program was completed in the early 1980s, leaving an important baseline data set for the study of land-cover change.

Each 1-degree block is split into sections to make files of manageable size; for example, the Albany, New York, 1-degree quadrangle has 15 sections. Data within blocks are stored in GIRAS format. Because the maps consist simply of land-cover polygons with their attributes, most of the data are nodes, lines, and links between the lines to create polygons—a loose topological structure. The map and section headers are followed by an arc records file, which consists of pointers (sequential links) into the next file, which contains the actual coordinates. The coordinates are in UTM and are truncated by rounding to the nearest 10 meters and by using the nearest 100,000 grid intersection.

Many of the USGS data sets are now available directly through the Internet by `ftp`. Chapter 6 gives details about how to search for these files. They are maintained on USGS servers usually listed alphabetically by the name of the map quadrangle concerned. This means that to retrieve the data set, the cartographer should both know what data are needed and what format the data are to be found in. Many major mapping and GIS software packages can now read these files directly, relieving the cartographer of the need to write computer programs to deal with the data formats.

The USGS also now distributes data on land-cover derived from classifications of NOAA's AVHRR (Advanced Very High Resolution Radiometer) measurements. These data are distributed by the EROS Data Center on CD-ROM, with a ground resolution of one kilometer. Biweekly composites showing a vegetation index for North America are

available, and efforts are under way to release a global AVHRR data set at this resolution.

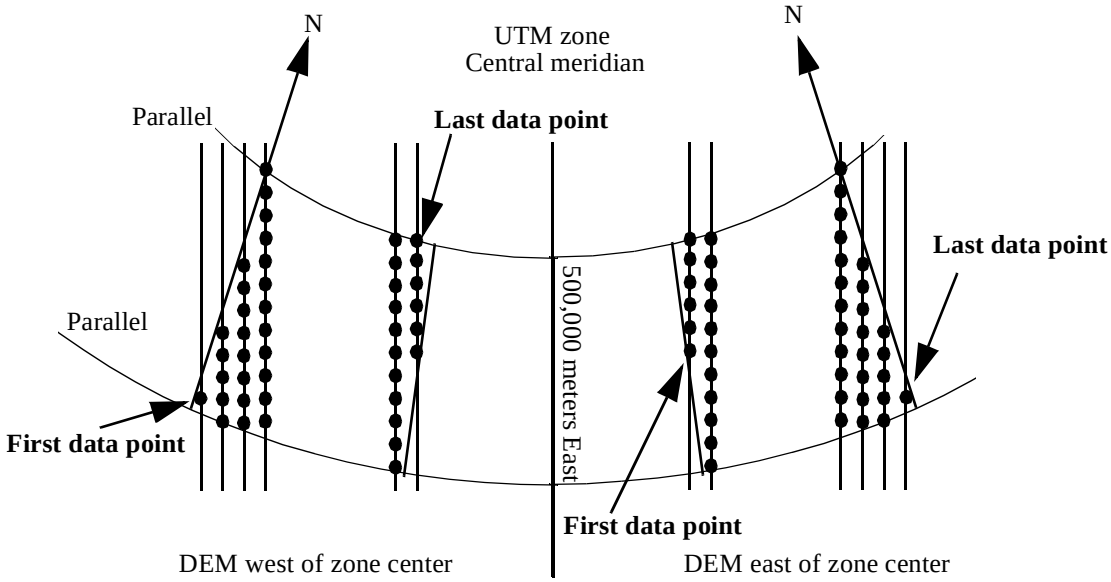

Figure 5.5 1:24,000 30 meter DEM format from the USGS.

5.4.2 The World Data Banks and the Digital Chart of the World

The CIA World Data Banks are widely distributed and used for many mapping purposes. They consist of decimal latitude and longitude pairs. World Data Bank I at a source map scale of 1:12,000,000 consisted of 100,000 *x, y* coordinates. World Data Bank II was digitized at 1:3,000,000 and has about six million coordinates. Neither database has a topological structure, although topologically structured versions exist in several proprietary formats. World Data Bank II has several topological inconsistencies. Each major coastline simply continues as a single line until it meets a significant point, such as a major national boundary, or closes on itself. The same is true of the political and hydrological data. The beginning of a new line is the only break from the sequence of binary-encoded pairs of latitude and longitude in degree-minute-second format. This structure, or lack of it, was the origin of the term "cartographic spaghetti." During this project—at a conference dinner table—the data were described as being "no more structured than spaghetti on a plate." Ever since, the entity-by-entity string structure has been alternatively known as "cartographic spaghetti."

An important new dataset, which largely replaces the World Data Banks, is the Digital Chart of the World (DCW). DCW is a digital version of the Defense Mapping Agency's Operational Navigation Chart (ONC) and Jet Navigation Charts (JNC) at a scale of 1:1,000,000. The data are digital vectors in string structure, formatted in a manner very

similar to the Vector Product Format (VPF) of the NATO DIGEST standard for digital data exchange covered in Chapter 6. The data cover some 14 layers, including coastlines, transportation, hydrology, contours, vegetation, inhabited places, and place names. The data are stored and distributed on four CD-ROMs, which include software for viewing the data files on IBM-PC compatible microcomputers. Distribution of the data is via the U.S. Geological Survey's Earth Science Data Centers. Information is available by calling 1-800-USA-MAPS in the United States and Canada.

DCW has some known problems that are emerging as the data are used for more and more applications. These include (1) problems of discontinuities at tile boundaries, because the data are tiled by latitude and longitude cells; (2) rings containing loops, such as an outer ring connecting to an included hole; and (3) rings that cross themselves. The size of the files, and the data being split across four CD-ROMs, means that queries and even simple draw operations on a microcomputer are extremely slow. Several efforts are under way to convert DCW to formats other than VPF, such as Arc/Info and DXF, and software vendors now offer their own software for DCW display.

The significance of the DCW for cartography cannot be understated. This single data set is likely to form the base layer for mapping around the world and in many ways is a fulfillment of the international efforts aimed at producing standardized world maps at this scale. The scale is particularly useful for integrating the broader satellite coverage such as AVHRR and some EOS data, and the detailed mapping coverages of the various nations of the world.

5.4.3 Formats at the Defense Mapping Agency

The Defense Mapping Agency (DMA) is the primary producer of digital cartographic data products for support of the U.S. military. The digital data form part of the Digital Landmass System (DLMS) and are made up of two parts, the Digital Feature Analysis Data (DFAD) and the Digital Terrain Elevation Data (DTED). Together, these data sets consist of about 25,000 magnetic tapes at 1,600 BPI (bits per inch, a tape data density), structured sequentially. This data set will eventually consist of 10^{19} bits of data. Overall, the DLMS contains information about terrain, landscape, and culture to support aircraft radar simulation, map and chart production, and navigation.

The DTED data contain elevation data sets in latitude, longitude, and elevation format that are used to generate elevations at 3-arc second intervals. These data are referenced to mean sea level as integer elevations rounded to the nearest meter. These data have been produced from contour maps and from stereo air photos, and have been enhanced with streambed and ridge information to preserve the topographic integrity.

The DFAD data contain descriptions of the land surface in terms of cultural and other features (forests, lakes, and such) stored as point, line, and area data. These data are lists of, or single latitude/longitude pairs and accompanying tables of attribute information. For example, a water tower would be stored in DFAD as a point and the attributes would include feature height and structure type, while a bridge may be recorded as a line segment. Linkage between the attributes and the cartographic data is maintained by header files.

The land features are coded at two levels. Level I data, with a resolution of about 152 meters (500 feet), contain large-area cultural information, such as surface material category, feature type, predominant height, structure density, and percentage roof and tree cover. Cultural features are digitized in a planimetric linear format, often called standard linear format (SLF). Standard linear format has been used as an internal digital cartographic data exchange format for the DMA.

Level II data are far more detailed, showing small area cultural information, and match the DTED data at 1-arc second resolution, about 30 meters (100 feet). These data are produced by manually digitizing large numbers of maps, charts, and air photos. The Level I data-bases will cover about 82 million square kilometers when complete. Level II data are digitized only for a limited number of areas of interest.

5.4.4 Formats at the National Ocean Service

Under the National Ocean Service (NOS) falls the Office of Coast and Geodetic Survey, and the National Oceanic and Atmospheric Administration (NOAA). These organizations produce two major types of cartographic products: nautical charts depicting hydrography and bathymetry and aeronautical charts which consist of visual flight charts and instrument flight charts. This division of the U.S. federal government conducts large-scale production of digital cartographic data, and some significant data sets are available.

The Nautical Charting Division of NOS offers digital shoreline data, and some aids to navigation. These data sets cover the shorelines of the coterminous United States, Puerto Rico, the U.S. Virgin Islands, and Hawaii, all to nautical chart accuracy standards. The data are available at small or large scale. The large-scale data have been digitized from the largest scale NOS nautical charts and plots of each part of the coastline, and they show by vectors the mean high-water mark. The shoreline vectors are continuous lines, but are broken where rivers flow into the ocean and where the shoreline cannot be precisely located. Each data set consists of between 100,000 and 300,000 points, and the data sets have been edge matched between data sets. Nineteen data sets cover the areas mentioned above, with approximately 20 more planned for Alaska. Three small-scale sets—East Coast, Gulf Coast, and West Coast—will be supplemented with an additional seven for the Great Lakes and Alaska.

The Nautical Charting Division of NOS has also embarked upon an ambitious program to convert all nautical charts to digital form as part of its ECDIS (Electronic Chart Display and Information System). As was the case at the USGS, these digital files were initially to be used to update and support standard printed map products. The newly evolved purpose is for these digital charts to be used by automated navigation systems for the piloting of ships and aircraft, as they now are. Efforts are under way to ensure that other nations of the world share data in a common format, thereby ensuring global standardized nautical charts for all the world's oceans.

The small-scale data are digitized from charts at 1:250,000 and show more generalized shorelines as a result. Records are 80 ASCII characters per record, written sequentially, with each point using one record. Each coordinate point is recorded as degrees, minutes, and seconds (to the nearest thousandth of a second) of latitude and longitude,

with full geodetic datum, scale, source, date, and feature code reference. In addition, separate files contain point data aids to navigation, such as the locations of wrecks, fixed and floating navigation aids, and obstructions. These data are stored in a similar format to the shoreline data and contain about 30,000 points per data set. Between the two data sets, all the data for the nautical charts are geocoded and distributed.

One final and notable digital data set is the global elevation data available from the National Geophysical Data Center, part of NOAA. These data are surface elevations and ocean depths, at 5 minutes of increment for latitude and longitude, digitized from navigational and aeronautical charts by the U.S. Navy and known as ETOPO5. The entire data set contains 2.3 million observations and is distributed both as the whole world and as segments by continent. The distribution medium is nine-track magnetic tape or floppy disk. In the recent past, the National Geophysical Data Center has released numerous digital map data sets on CD-ROM, most recently including detailed bathymetry of the ocean and land-surface topography as well as geodetic and magnetic data for the earth's surface.

5.4.5 Digital Image Formats

Digital imagery is increasingly an important source of data for integration with digital vector maps in GIS. Two major types of digital imagery are of cartographic importance: scanned air photos and remotely sensed satellite and aircraft data. Numerous government agencies provide image data, including the National Aeronautics and Space Administration (NASA), the U.S. Geological Survey, and the National Ocean Service. Similarly, private companies distribute data from commercial land remote sensing programs, including EOSAT and SPOT.

NASA owns and operates numerous instruments and satellites for the collection of Earth data that are suitable for mapping. A very large amount of data from aircraft and spacecraft for Earth and the planets and their moons are available for cartographic purposes. The bulk of the data is in the public domain, and is available at the cost of reproduction to users. NASA maintains a central directory system for archiving and searching its data, which are scattered around the major NASA centers around the United States. The NASA Master Directory is covered in Chapter 6, because it is network accessible.

As part of the Mission to Planet Earth, NASA is building the Earth Observation System (EOS) (see Price et al., 1994). EOS data, from a very large array of instruments including those suitable for land-surface mapping, will be distributed through the NASA EOSDIS (EOS Data and Information System). The system will consist of a number of Distributed Active Archive Centers (DAACs), located mostly at NASA centers. The Land Surface Processes DAAC is under construction at the EROS Data Center in Sioux Falls, South Dakota. DAACs will be publicly accessible through the Internet and will support browsing and file transfer to geographic information scientists. An immense reserve of remotely sensed data will soon be available for mapping from these centers.

The USGS operates the central archive in the United States for the Landsat and AVHRR (Advanced Very High Resolution Radiometer) data provided by U.S. government mapping satellites.The data is stored on magnetic reel and cartridge tape at the USGS National Mapping Division's EROS Data Center in Sioux Falls, South Dakota.

All public domain Landsat data, and a large variety of other image format data, are available from this site. The EROS Data Center serves as the distribution point for many other types of digital data, such as DEMs. AVHRR data are available on an hourly basis, with 4- and 1-kilometer ground resolution. As such, they are used for intermediate and small-scale mapping of global, continental, and regional areas, including the United States. Sample data and useful overlays such as topography are available on CD-ROM.

The USGS also provides two other sets of digital image data of interest to cartography. These are the SLAR (Side Looking Airborne Radar) images, corresponding to 1:250,000 coverage of the United States, as part of an ongoing radar mapping program, and the new Digital Orthophotographic Quadrangle (DOQs).

The DOQ is likely to become a primary source of map information in the United States, especially for map updating and when a high-resolution map reference base is necessary. Distribution is by CD-ROM in quadrangle format, with four 1:12,000-equivalent quarter quads making up areas consistent with the USGS 7.5-minute series. The files are JPEG compressed gray scale (single band) images of about 50 megabytes apiece, with a ground equivalent resolution of 1 meter. The USGS, with assistance from the U.S. Department of Agriculture, plans to map the whole country in the DOQ format and to update the coverage on a five-year cycle.

Finally, an important source of up-to-date information on weather and atmospheric conditions is the GOES weather satellite data. Although this data are of very poor spatial resolution and is grossly distorted by Earth's curvature, its ready availability has made it popular with navigators and weather forecasters, including television stations. The data are network accessible; access is discussed in Chapter 6.

5.4.6 Industry "Standard" Formats

Several industry standards have emerged to satisfy the immediate data transfer needs in computer cartography. Although none of these data formats is accepted by national or international organizations as transfer standards, proprietary standards are nevertheless extremely common in the work environment.

At the initial level, many software packages maintain their own format and file structures for data in use by the software. Arc/Info, for example, uses protected file formats and structures the files according to projects, geometry, and relationships between the data. All files containing data for the INFO relational database manager, for example, must be in a separate directory under the project directory. Similarly, most commercial systems use proprietary formats to optimize storage efficiency, increase storage, or optimize a system around a particular piece of hardware. Most packages that use proprietary formats contain modules that read (and sometimes write) many of the formats covered in this chapter. Without this capability, users would be frustrated because they would be unable to bring in new cartographic data.

Some proprietary formats have emerged as common to many systems and therefore serve as *de facto* exchange standards (Kay and Levine, 1992). None of these is optimized for cartographic applications, such as allowing the exchange of cartographic geometry, topology, and attributes. Nevertheless, many systems allow exchange of attributes from

statistical packages such as SAS, from spreadsheets, or from database packages, thus allowing the reassembling of the information after the transfer. Many systems do not handle the transfer of topology at all, requiring its reconstruction when the map is transported between computers.

For vector data, the three most common industry standards are AutoCad DXF, PostScript, and HPGL. DXF, for AutoDesk's Digital eXchange Format, is broadly used in the computer-assisted drafting and design field. The format consists of a large number of default ASCII text labels, accompanied on the following line by numerical values for these fields. At the core is a list of *x* and *y* values defining the drawing, in this case forming the lines on a map. DXF supports structuring by layers, a concept familiar in GIS, but in a graphic rather than a topological sense. Both PostScript and HPGL are really plotter formats, designed to allow hardware devices (laser-jet printers using the Adobe software and fonts, and Hewlett Packard printers and plotters) to translate *x* and *y* increments into firmware plot commands. Typically, these formats involve a base level with ASCII codes and coordinates corresponding to move and draw commands and other structured commands which establish the plotter size, pen colors, text font, and size and so forth.

For raster data, a particularly rich set of industry standards has developed. Among the leading formats are Targa, PICT, TIFF, GIF, PCX, JPEG, and Encapsulated PostScript. Many scanners, which generate a raster format, and bitmap editors such as PC Paintbrush, SuperPaint, Adobe Photoshop, and IslandPaint, import and export combinations of these formats. A few packages can import and export both raster and vector formats, such as CorelDraw! and Adobe Illustrator. In addition, several packages are available to translate between these formats, such as Freedom of the Press and Image Alchemy.

Each of these formats is similar, with the only exceptions being those using compression schemes (in this case GIF and JPEG). Many of the formats work across bit depths, so that a 24-bit image will be displayable in binary, for example. Targa, PCX, and TIFF files are similarly structured, in that they consist of a single file containing a file header, a color map, and a set of color indexes. The color map is a set of Red, Green, and Blue (RGB) intensities onto which the data values are mapped. Thus color index 10, for example, may consist of the color pure red (red = 255, green = 0, blue = 0). As the image grid is extracted from the file, all values of 10 are then given these RGB values for display. In systems that advertise a selection of 256 possible colors from a "palette" of over 16 million, the 16 million comes from being able to index $256 \times 256 \times 256$ colors, but only 256 at a time since the data array itself is only eight bits deep.

JPEG and GIF formats implement compression strategies, of which the most sophisticated is that used by JPEG. JPEG is able to compress files with some degree of redundancy in their underlying data many times, for example 50:1. JPEG also allows partial recompression,that is, lossy compression, in which some of the image data are lost to further compress the data. This is a significant problem for many cartographic applications, where loss-less compression is preferred. The lineage of these file structures varies. Targa files were designed to support a particular graphics display board for microcomputers. GIF files were designed for shipping images over networks, in particular the CompuServe system.

PCX is proprietary to Zsoft, but is supported (along with the BMP Windows bitmap format) in the Paintbrush drawing accessory program which comes free with Microsoft's

Windows operating system for IBM-PC compatible microcomputers. TIFF formats have been more popular for use on Apple Macintosh computers, in software packages such as Adobe Illustrator, and for scanners.

For specific needs, these formats are often more than adequate. Most major software packages for computer cartography will deal with one or more of these formats. Rarely are these internal or programming formats, however, but instead they are invoked when data are to be stored as external binary or ASCII files on disk. Because none of the formats is a national or international standard, their future is not guaranteed. Archiving data for long periods in these formats is likely to entail a significant risk.

To counteract the problems associated with long-term storage and exchange of data over noncompatible systems, several formal standards efforts have been initiated and in a couple of cases completed. These standard formats represent the best possibility for digital cartographic data to be both broadly distributed and exchanged between cartographers and geographic information scientists. The standards efforts are covered in detail in the Chapter 6, and one standard, the Spatial Data Transfer Standard, is used in the development of a set of C language data formats, models, and structures for use in computer programming.

5.5 REFERENCES

Defense Mapping Agency (1985a). *Standard Linear Format.* Washington, DC: U.S. Government Printing Office.

Defense Mapping Agency (1985b). *Feature Attribute Coding Standard.* Washington, D.C: U.S. Government Printing Office.

Department of the Interior, U.S. Geological Survey (1985a). *Digital Line Graphs from 1:100,000-Scale Maps, Data Users Guide.* National Mapping Program Technical Instructions, Data Users Guide 2, Reston, VA.

Department of the Interior, U.S. Geological Survey (1985b). *Geographic Names Information System, Data Users Guide.* National Mapping Program Technical Instructions, Data Users Guide 6, Reston, VA.

Department of the Interior, U.S. Geological Survey (1986a). *Land-use and Land-cover Digital Data from 1:250,000- and 100,000-Scale Maps, Data Users Guide.* National Mapping Program Technical Instructions, Data Users Guide 4, Reston, VA.

Department of the Interior, U.S. Geological Survey (1986b). *Digital Line Graphs from 1:24,000-Scale Maps, Data Users Guide.* National Mapping Program Technical Instructions, Data Users Guide 1, Reston, VA.

Department of the Interior, U.S. Geological Survey (1987a). *Digital Line Graphs from 1:2,000,000-Scale Maps, Data Users Guide.* National Mapping Program Technical Instructions, Data Users Guide 3, Reston, VA.

Department of the Interior, U.S. Geological Survey (1987b). *Digital Elevation Models, Data Users Guide.* National Mapping Program Technical Instructions, Data Users Guide 5, Reston, VA.

Guptill, S. C., ed. (1990). *An Enhanced Digital Line Graph Design*. Department of the Interior, U.S. Geological Survey Circular 1048. Washington, DC: U.S. Government Printing Office.

Held, G. (1983). *Data Compression: Techniques and Applications, Hardware and Software Considerations*. New York: Wiley.

Kay, D. C., and J. R. Levine, (1992). *Graphics File Formats.* Blue Ridge Summit, PA: Windcrest/McGraw-Hill.

Price, P. D., et al. (1994). "Earth Science Data for All: EOS and the EOS Data System." *Photogrammetric Engineering and Remote Sensing*, vol. 60, no. 3, pp. 277–285.