# CHAPTER 3: GEOGRAPHYCAL INFORMATION SYSTEM (GIS)

## What is GIS – Definition?

Geographic Information System (GIS) is a computer system build to capture, store, manipulate, analyze, manage and display all kinds of spatial or geographical data. GIS applications are tools that allow end users to perform spatial query, analysis, edit spatial data and create hard copy maps. In simple way GIS can be define as an image that is referenced to the earth or has x and y coordinate and it‘s attribute values are stored in the table. These x and y coordinates are based on different projection system and there are various types of projection system. Most of the time GIS is used to create maps and to print. To perform the basic task in GIS, layers are combined, edited and designed.

GIS can be used to solve the location based question such as ―What is located here‖ or Where to find particular features? GIS User can retrieve the value from the map, such as how much is the forest area on the land use map. This is done using the query builder tool. Next important features of the GIS is the capability to combine different layers to show new information. For example, you can combine elevation data, river data, land use data and many more to show information about the landscape of the area. From map you can tell where is high lands or where is the best place to build house, which has the river view . GIS helps to find new information.

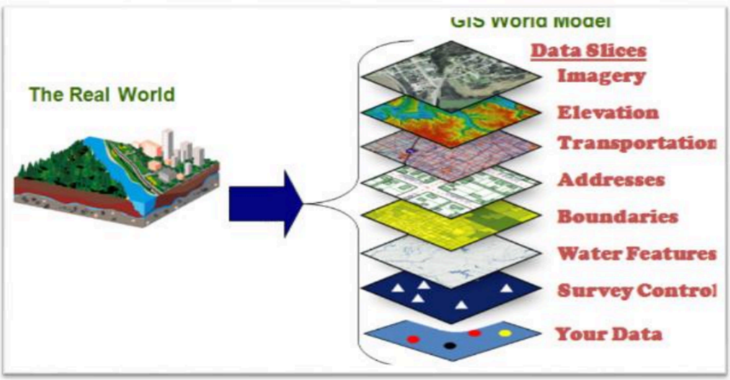



Figure: Gis Model

## How GIS Works:

- Visualizing Data: The geographic data that is stored in the databases are displayed in the GIS software.

- Combining Data: Layers are combined to form a maps of desire.
- The Query: To search the value in the layer or making a geographic queries.

Definition by others:

A geographic information system (GIS) lets us visualize, question, analyze, and interpret data to understand relationships, patterns, and trends. (ESRI)

In the strictest sense, a GIS is a computer system capable of assembling, storing, manipulating, and displaying geographically referenced information (that is data identified according to their locations). (USGS)

Advantage of GIS:

- Better decision made by government people
- Improve decision making with the help of layered information
- Citizen engagement due to better system
- Help to identify communities that is under risk or lacking infrastructure
- Helps in identifying criminology matters
- Better management of natural resources
- Better communication during emergency situation
- Cost savings due to better decision
- Finding different kinds of trends within the community
- Planning the demographic changes

## History of GIS:

Modern GIS has seen series of development. GIS has evolved with the computer system. Here are the brief events that has happened for the development of the GIS system.

Year 1854 – The term GIS that used scientific method to create maps was used by John Snow in 1854. He used points on London residential map to plot outbreak of Cholera.

Year 1960 – Modern computerized GIS system began in year 1960.

Year 1962 – Dr. Roger Tomlinson created and developed Canadian Geographic Information System (CGIS) to store, analyze and manipulate data that was collected for the Canada Land Inventory (CLI). This software had the capacity to overlay, measurement and digitizing (converting scan hardcopy map to digital data). It is never provided in commercial format but Dr. Tomlinson is the father of GIS.

Year 1980 – This period saw rise of commercial GIS software‘s like M&S Computing, Environmental Systems Research Institute (ESRI) and Computer Aided Resource Information System (CARIS). These all software were similar to CGIS with more functionality and user-friendliness. Among all the above the most popular today is ESRI products like ArcGIS, ArcView which hold almost 80 % of global market.

## Component of GIS:

Hardware: Hardware is the physical component of the computer and GIS runs on it. Hardware may be hard disk, processor, motherboard and so on. All these hardware work together to function as a computer. GIS software run on these hardware. Computer can be standalone called desktop or server based. GIS can run on both of them.

Software: GIS Software provides tools and functions to input and store spatial data or geographic data. It provides tool to perform geographic query, run analysis model and display geographic data in the map form. GIS software uses Relation Database Management System (RDBMS) to store the geographic data. Software talks with the database to perform geographic query.

Data: Data are the fuel for the GIS and the most important and expensive component. Geographic data are the combination of physical features and it‘s information which is stored in the tables. These tables are maintained by the RDBMS. The process of capturing the geographic data are called digitization which is the most tedious job. It is the process of converting scanned hardcopy maps into the digital format. Digitization is done by tracing the lines along the geographic features for example to capture a building you will trace around the building on the image.

People: People are the user of the GIS system.
People use all above three component to run a GIS system. Today‘s computer are fast and user friendly which makes it easy to perform geographic queries, analysis and displaying maps.
Today everybody uses GIS to perform their daily job.

## Area of GIS Applications

Major areas of GIS application can be grouped into five categories as follows.

Facilities Management: Large scale and precise maps and network analysis are used mainly for utility management. AM/FM is frequently used in this area.

Environment and Natural Resources Management: Medium or small scale maps and overlay techniques in combination with aerial photographs and satellite images are used for management of natural resources and environmental impact analysis.

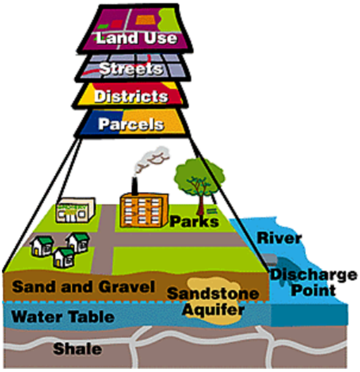
Street Network: Large or medium scale maps and spatial analysis are used for vehicle routing, locating house and streets etc.

Planning and Engineering: Large or medium scale maps and engineering models are used mainly in civil engineering.

Land Information System: Large scale cadastre maps or land parcel maps and spatial analysis are used for cadastre administration, taxation etc.

## GIS Data Models

A GIS stores information about the world as a collection of thematic layers that can be linked together by geography. This simple but extremely powerful and versatile concept has proven invaluable for solving many real-world problems from tracking delivery vehicles, to recording details of planning applications, to modeling global atmospheric circulation. The thematic layer approach allows us to organize the complexity of the real world into a simple representation to help facilitate our understanding of natural relationships.




The thematic layer approach allows us to organize the complexity of the real world

## GIS Data Types

The basic data type in a GIS reflects traditional data found on a map. Accordingly, GIS technology utilizes two basic types of data. These are:

- Spatial data: describes the absolute and relative location of geographic features.
- Attribute data: describes characteristics of the spatial features. These characteristics can be quantitative and/or qualitative in nature. Attribute data is often referred to as tabular data.

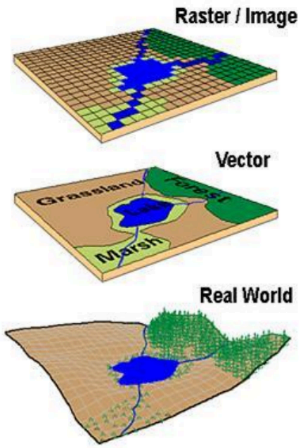
The coordinate location of a forestry stand would be spatial data, while the characteristics of that forestry stand, e.g. cover group, dominant species, crown closure, height, etc., would be attribute data. Other data types, in particular image and multimedia data, are becoming more prevalent with changing technology. Depending on the specific content of the data, image data may be considered either spatial, e.g. photographs, animation, movies, etc., or attribute, e.g. sound, descriptions, narration's, etc.

# Spatial Data Models

Traditionally spatial data has been stored and presented in the form of a map. Three basic types of spatial data models have evolved for storing geographic data digitally. These are referred to as:

1. Vector;
2. Raster;
3. Image.

The following diagram reflects the two primary spatial data encoding techniques. These are vector and raster. Image data utilizes techniques very similar to raster data, however typically lacks the internal formats required for analysis and modeling of the data. Images reflect pictures or photographs of the landscape.




Representation of the real world and showing differences in how a vector and a raster GIS will represent this real world.

<u>Vector Data Formats</u>

All spatial data models are approaches for storing the spatial location of geographic features in a database. Vector storage implies the use of vectors (directional lines) to represent a geographic feature. Vector data is characterized by the use of sequential points or vertices to define a linear segment. Each vertex consists of an X coordinate and a Y coordinate.

Vector lines are often referred to as arcs and consist of a string of vertices terminated by a node. A node is defined as a vertex that starts or ends an arc segment. Point features are defined by one coordinate pair, a vertex. Polygonal features are defined by a set of closed coordinate pairs. In vector representation, the storage of the vertices for each feature is important, as well as the connectivity between features, e.g. the sharing of common vertices where features connect.

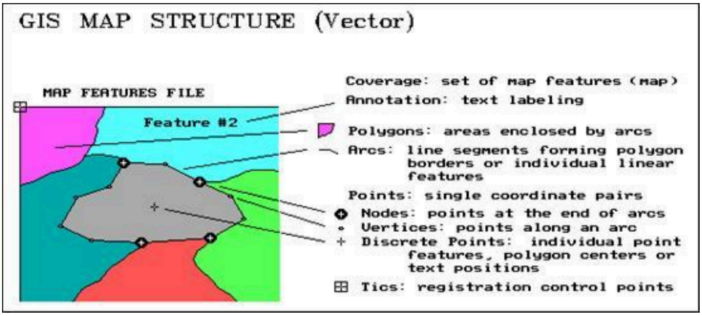
Vector Data: There are three types of vector data, points, lines and polygons. These data are created by digitizing the base data. They store information in x, y coordinates. Vectors models are used to store data which have discrete boundaries like country borders, land parcels and roads.

Several different vector data models exist, however only two are commonly used in GIS data storage.

The most popular method of retaining spatial relationships among features is to explicitly record adjacency information in what is known as the topologic data model. Topology is a mathematical concept that has its basis in the principles of feature adjacency and connectivity.

The topologic data structure is often referred to as an intelligent data structure because spatial relationships between geographic features are easily derived when using them. Primarily for this reason the topologic model is the dominant vector data structure currently used in GIS technology. Many of the complex data analysis functions cannot effectively be undertaken without a topologic vector data structure. Topology is reviewed in greater detail later on in the book.

The secondary vector data structure that is common among GIS software is the computer-aided drafting (CAD) data structure. This structure consists of listing elements, not features, defined by strings of vertices, to define geographic features, e.g. points, lines, or areas. There is considerable redundancy with this data model since the boundary segment between two polygons can be stored twice, once for each feature. The CAD structure emerged from the development of computer graphics systems without specific considerations of processing geographic features. Accordingly, since features, e.g. polygons, are self-contained and independent, questions about the adjacency of features can be difficult to answer. The CAD vector model lacks the definition of spatial relationships between features that is defined by the topologic data model.




GIS MAP Structure - VECTOR systems (Adapted from Berry)

Raster Data Formats

Raster data models incorporate the use of a grid-cell data structure where the geographic area is divided into cells identified by row and column. This data structure is commonly called raster. While the term raster implies a regularly spaced grid other tessellated data structures do exist in grid based GIS systems. In particular, the quadtree data structure has found some acceptance as an alternative raster data model.

Raster Data: Raster data store information of features in cell based manner. Satellite images, photogrammetry and scanned maps are all raster based data. Raster model are used to store data which varies continuously as in aerial photography, a satellite image or elevation values (DEM-Digital Elevation Model).

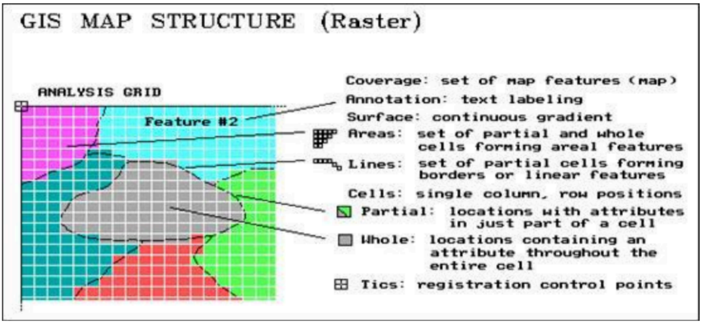
The size of cells in a tessellated data structure is selected on the basis of the data accuracy and the resolution needed by the user. There is no explicit coding of geographic coordinates required since that is implicit in the layout of the cells. A raster data structure is in fact a matrix where any coordinate can be quickly calculated if the origin point is known, and the size of the grid cells is known. Since grid-cells can be handled as two-dimensional arrays in computer encoding many analytical operations are easy to program. This makes tessellated data structures a popular choice for many GIS software. Topology is not a relevant concept with tessellated structures since adjacency and connectivity are implicit in the location of a particular cell in the data matrix.

Several tessellated data structures exist, however only two are commonly used in GIS's. The most popular cell structure is the regularly spaced matrix or raster structure. This data structure involves a division of spatial data into regularly spaced cells. Each cell is of the same shape and size. Squares are most commonly utilized.

Since geographic data is rarely distinguished by regularly spaced shapes, cells must be classified as to the most common attribute for the cell. The problem of determining the proper resolution for a particular data layer can be a concern. If one selects too coarse a cell size then data may be overly generalized. If one selects too fine a cell size then too many cells may be created resulting in a large data volume, slower processing times, and a more cumbersome data set. As well, one can imply accuracy greater than that of the original data capture process and this may result in some erroneous results during analysis.

As well, since most data is captured in a vector format, e.g. digitizing, data must be converted to the raster data structure. This is called vector-raster conversion. Most GIS software allows the user to define the raster grid (cell) size for vector-raster conversion. It is imperative that the original scale, e.g. accuracy, of the data be known prior to conversion. The accuracy of the data, often referred to as the resolution, should determine the cell size of the output raster map during conversion.

Most raster based GIS software requires that the raster cell contain only a single discrete value. Accordingly, a data layer, e.g. forest inventory stands, may be broken down into a series of raster maps, each representing an attribute type, e.g. a species map, a height map, a density map, etc. These are often referred to as one attribute maps. This is in contrast to most conventional vector data models that maintain data as multiple attribute maps, e.g. forest inventory polygons linked to a database table containing all attributes as columns. This basic distinction of raster data storage provides the foundation for quantitative analysis techniques. This is often referred to as raster or map algebra. The use of raster data structures allow for sophisticated mathematical modelling processes while vector based systems are often constrained by the capabilities and language of a relational DBMS.




GIS MAP Structure - RASTER systems (Adapted from Berry)

This difference is the major distinguishing factor between vector and raster based GIS software. It is also important to understand that the selection of a particular data structure can provide advantages during the analysis stage. For example, the vector data model does not handle continuous data, e.g. elevation, very well while the raster data model is more ideally suited for this type of analysis. Accordingly, the raster structure does not handle linear data analysis, e.g. shortest path, very well while vector systems do. It is important for the user to understand that there are certain advantages and disadvantages to each data model.

The selection of a particular data model, vector or raster, is dependent on the source and type of data, as well as the intended use of the data. Certain analytical procedures require raster data while others are better suited to vector data.

Image Data

Image data is most often used to represent graphic or pictorial data. The term image inherently reflects a graphic representation, and in the GIS world, differs significantly from raster data. Most often, image data is used to store remotely sensed imagery, e.g. satellite scenes or orthophotos, or ancillary graphics such as photographs, scanned plan documents, etc. Image data is typically used in GIS systems as background display data (if the image has been rectified and georeferenced); or as a graphic attribute. Remote sensing software makes use of image data for image classification and processing. Typically, this data must be converted into a raster format (and perhaps vector) to be used analytically with the GIS.

Image data is typically stored in a variety of de facto industry standard proprietary formats. These often reflect the most popular image processing systems. Other graphic image formats, such as TIFF, GIF, PCX, etc., are used to store ancillary image data. Most GIS software will read such formats and allow you to display this data.

Image data is most often used for remotely sensed imagery such as satellite imagery or digital orthophotos.

Vector and Raster - advantages and disadvantages

There are several advantages and disadvantages for using either the vector or raster data model to store spatial data. These are summarized below.

| Vector Data | |
|---|---|
| Advantages :<br>1)Data can be represented at its original resolution and form without generalization.<br>2)Graphic output is usually more aesthetically pleasing (traditional cartographic representation);<br>3)Since most data, e.g. hard copy maps, is in vector form no data conversion is required.<br>4)Accurate geographic location of data is maintained.<br>5)Allows for efficient encoding of | Disadvantages:<br>1) The location of each vertex needs to be stored explicitly.<br>2) For effective analysis, vector data must be converted into a topological structure. This is often processing intensive and usually requires extensive data cleaning. As well, topology is static, and any updating or editing of the vector data requires re-building of the topology.<br>3) Algorithms for manipulative and analysis functions are complex and may be processing intensive. Often, this inherently limits the functionality for large data sets, e.g. a large |

| | |
|---|---|
| topology, and as a result more efficient operations that require topological information, e.g. proximity, network analysis. | number of features.<br>4) Continuous data, such as elevation data, is not effectively represented in vector form. Usually substantial data generalization or interpolation is required for these data layers.<br>5) Spatial analysis and filtering within polygons is impossible |
| Raster Data | |
| Advantages :<br>1) The geographic location of each cell is implied by its position in the cell matrix. Accordingly, other than an origin point, e.g. bottom left corner, no geographic coordinates are stored.<br>2) Due to the nature of the data storage technique data analysis is usually easy to program and quick to perform.<br>3) The inherent nature of raster maps, e.g. one attribute maps, is ideally suited for mathematical modeling and quantitative analysis.<br>4) Discrete data, e.g. forestry stands, is accommodated equally well as continuous data, e.g. elevation data, and facilitates the integrating of the two data types.<br>5) Grid-cell systems are very compatible with raster-based output devices, e.g. electrostatic plotters, graphic terminals. | Disadvantages:<br>1) The cell size determines the resolution at which the data is represented.<br>2) It is especially difficult to adequately represent linear features depending on the cell resolution. Accordingly, network linkages are difficult to establish.<br>3) Processing of associated attribute data may be cumbersome if large amounts of data exists. Raster maps inherently reflect only one attribute or characteristic for an area.<br>4) Since most input data is in vector form, data must undergo vector-to-raster conversion. Besides increased processing requirements this may introduce data integrity concerns due to generalization and choice of inappropriate cell size.<br>5) Most output maps from grid-cell systems do not conform to high-quality cartographic needs. |

It is often difficult to compare or rate GIS software that use different data models. Some personal computer (PC) packages utilize vector structures for data input, editing, and display but convert to raster structures for any analysis. Other more comprehensive GIS offerings provide both integrated raster and vector analysis techniques. They allow users to select the data structure appropriate for the analysis requirements. Integrated raster and vector processing capabilities are most desirable and provide the greatest flexibility for data manipulation and analysis.

## Attribute Data Models

A separate data model is used to store and maintain attribute data for GIS software. These data models may exist internally within the GIS software, or may be reflected in external commercial Database Management Software (DBMS). A variety of different data models exist for the storage and management of attribute data. The most common are:

- Tabular
- Hierarchial
- Network

- Relational
- Object Oriented

The tabular model is the manner in which most early GIS software packages stored their attribute data. The next three models are those most commonly implemented in database management systems (DBMS). The object oriented is newer but rapidly gaining in popularity for some applications. A brief review of each model is provided.

Tabular Model

The simple tabular model stores attribute data as sequential data files with fixed formats (or comma delimited for ASCII data), for the location of attribute values in a predefined record structure. This type of data model is outdated in the GIS arena. It lacks any method of checking data integrity, as well as being inefficient with respect to data storage, e.g. limited indexing capability for attributes or records, etc.

Hierarchical Model

The hierarchical database organizes data in a tree structure. Data is structured downward in a hierarchy of tables. Any level in the hierarchy can have unlimited children, but any child can have only one parent. Hierarchial DBMS have not gained any noticeable acceptance for use within GIS. They are oriented for data sets that are very stable, where primary relationships among the data change infrequently or never at all. Also, the limitation on the number of parents that an element may have is not always conducive to actual geographic phenomenon.

Network Model

The network database organizes data in a network or plex structure. Any column in a plex structure can be linked to any other. Like a tree structure, a plex structure can be described in terms of parents and children. This model allows for children to have more than one parent.

Network DBMS have not found much more acceptance in GIS than the hierarchical DBMS. They have the same flexibility limitations as hierarchical databases; however, the more powerful structure for representing data relationships allows a more realistic modelling of geographic phenomenon. However, network databases tend to become overly complex too easily. In this regard it is easy to lose control and understanding of the relationships between elements.

Relational Model

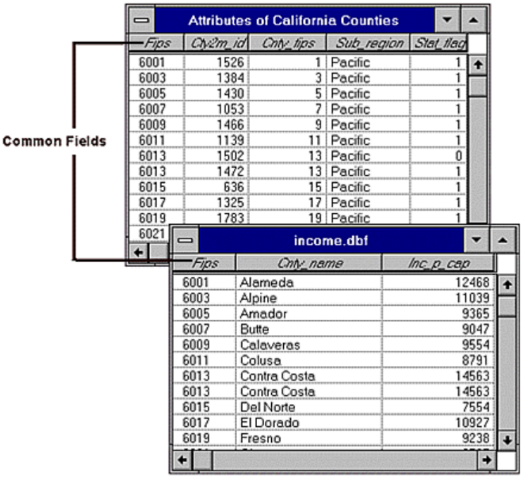
The relational database organizes data in tables. Each table, is identified by a unique table name, and is organized by rows and columns. Each column within a table also has a unique name. Columns store the values for a specific attribute, e.g. cover group, tree height. Rows represent one record in the table. In a GIS each row is usually linked to a separate spatial feature, e.g. a forestry stand. Accordingly, each row would be comprised of several columns, each column containing a specific value for that geographic feature. The following figure presents a sample table for forest inventory features. This table has 4 rows and 5 columns. The forest stand number would be the label for the spatial feature as well as the primary key for the database table. This serves as the linkage between the spatial definition of the feature and the attribute data for the feature.

| UNIQUE STAND NUMBER | DOMINANT COVER GROUP | AVG. TREE HEIGHT | STAND SITE INDEX | STAND AGE |
|---|---|---|---|---|
| 001 | DEC | 3 | G | 100 |
| 002 | DEC-CON | 4 | M | 80 |
| 003 | DEC-CON | 4 | M | 60 |
| 004 | CON | 4 | G | 120 |

Data is often stored in several tables. Tables can be joined or referenced to each other by common columns (relational fields). Usually the common column is an identification number for a selected geographic feature, e.g. a forestry stand polygon number. This identification number acts as the primary key for the table. The ability to join tables through use of a common column is the essence of the relational model. Such relational joins are usually ad hoc in nature and form the basis of for querying in a relational GIS product. Unlike the other previously discussed database types, relationships are implicit in the character of the data as opposed to explicit characteristics of the database set up.

The relational database model is the most widely accepted for managing the attributes of geographic data.

There are many different designs of DBMSs, but in GIS the relational design has been the most useful. In the relational design, data are stored conceptually as a collection of tables. Common fields in different tables are used to link them together. This surprisingly simple design has been so widely used primarily because of its flexibility and very wide deployment in applications both within and without GIS.




| Fips | Cty2m_id | Cnty_fips | Sub_region | Stat_flag |
|---|---|---|---|---|
| 6001 | 1526 | 1 | Pacific | 1 |
| 6003 | 1384 | 3 | Pacific | 1 |
| 6005 | 1430 | 5 | Pacific | 1 |
| 6007 | 1053 | 7 | Pacific | 1 |
| 6009 | 1466 | 9 | Pacific | 1 |
| 6011 | 1139 | 11 | Pacific | 1 |
| 6013 | 1502 | 13 | Pacific | 0 |
| 6013 | 1472 | 13 | Pacific | 1 |
| 6015 | 636 | 15 | Pacific | 1 |
| 6017 | 1325 | 17 | Pacific | 1 |
| 6019 | 1783 | 19 | Pacific | 1 |
| 6021 | | | | |

| Fips | Cnty_name | Inc_p_cap |
|---|---|---|
| 6001 | Alameda | 12468 |
| 6003 | Alpine | 11039 |
| 6005 | Amador | 9365 |
| 6007 | Butte | 9047 |
| 6009 | Calaveras | 9554 |
| 6011 | Colusa | 8791 |
| 6013 | Contra Costa | 14563 |
| 6013 | Contra Costa | 14563 |
| 6015 | Del Norte | 7554 |
| 6017 | El Dorado | 10927 |
| 6019 | Fresno | 9238 |

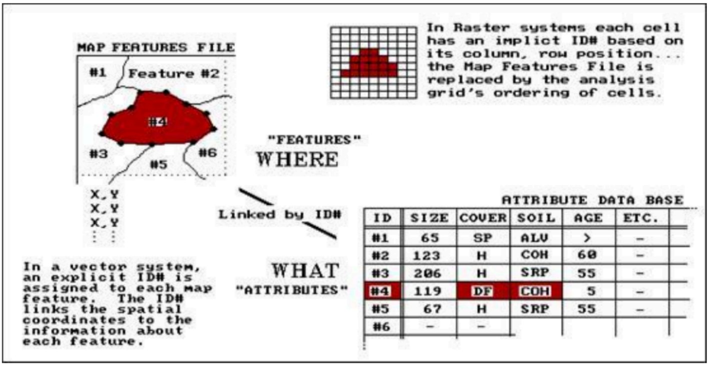
In the relational design, data are stored conceptually as a collection of tables. Common fields in different tables are used to link them together.

In fact, most GIS software provides an internal relational data model, as well as support for commercial off-the-shelf (COTS) relational DBMS'. COTS DBMS' are referred to as external DBMS'. This approach supports both users with small data sets, where an internal data model is sufficient, and customers with larger data sets who utilize a DBMS for other corporate data storage requirements. With an external DBMS the GIS software can simply connect to the database, and the user can make use of the inherent capabilities of the DBMS. External DBMS' tend to have much more extensive querying and data integrity capabilities than the GIS' internal relational model. The emergence and use of the external DBMS is a trend that has resulted in the proliferation of GIS technology into more traditional data processing environments.

The relational DBMS is attractive because of its:
1) simplicity in organization and data modelling.
2) flexibility - data can be manipulated in an ad hoc manner by joining tables.
3) efficiency of storage - by the proper design of data tables redundant data can be minimized; and
4) the non-procedural nature - queries on a relational database do not need to take into account the internal organization of the data.

The relational DBMS has emerged as the dominant commercial data management tool in GIS implementation and application.

The following diagram illustrates the basic linkage between a vector spatial data (topologic model) and attributes maintained in a relational database file.




| ID | SIZE | COVER | SOIL | AGE | ETC. |
|---|---|---|---|---|---|
| #1 | 65 | SP | ALU | > | - |
| #2 | 123 | H | COH | 60 | - |
| #3 | 206 | H | SRP | 55 | - |
| #4 | 119 | DF | COH | 5 | - |
| #5 | 67 | H | SRP | 55 | - |
| #6 | - | - | | | |

Basic linkages between a vector spatial data (topologic model) and attributes maintained in a relational database file (From Berry)

Object-Oriented Model

The object-oriented database model manages data through objects. An object is a collection of data elements and operations that together are considered a single entity. The object-oriented database is a relatively new model. This approach has the attraction that querying is very natural, as features can be bundled together with attributes at the database administrator's discretion. To date, only a few GIS packages are promoting the use of this attribute data model. However, initial impressions indicate that this approach may hold many operational benefits with respect to geographic data processing. Fulfilment of this promise with a commercial GIS product remains to be seen.

## Sources of Data

As previously identified, two types of data are input into a GIS, spatial and attribute. The data input process is the operation of encoding both types of data into the GIS database formats.

The creation of a clean digital database is the most important and time consuming task upon which the usefulness of the GIS depends. The establishment and maintenance of a robust spatial database is the cornerstone of a successful GIS implementation.

As well, the digital data is the most expensive part of the GIS. Yet often, not enough attention is given to the quality of the data or the processes by which they are prepared for automation.

The general consensus among the GIS community is that 60 to 80 % of the cost incurred during implementation of GIS technology lies in data acquisition, data compilation and database development.

A wide variety of data sources exist for both spatial and attribute data. The most common general sources for spatial data are:

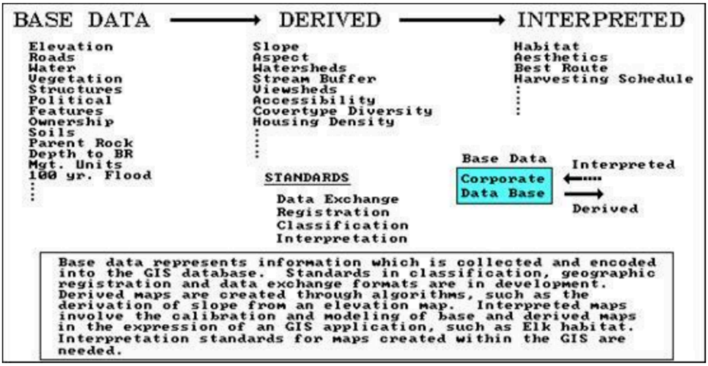
- hard copy maps;
- aerial photographs;
- remotely-sensed imagery;
- point data samples from surveys; and
- existing digital data files.

Existing hard copy maps, e.g. sometimes referred to as analogue maps, provide the most popular source for any GIS project.

Potential users should be aware that while there are many private sector firms specializing in providing digital data, federal, provincial and state government agencies are an excellent source of data. Because of the large costs associated with data capture and input, government departments are often the only agencies with financial resources and manpower funding to invest in data compilation. British Columbia and Alberta government agencies are good examples. Both provincial governments have defined and implemented province wide coverage of digital base map data at varying map scales, e.g. 1:20,000 and 1:250,000. As well, the provincial forestry agencies also provide thematic forest inventory data in digital format. Federal agencies are also often a good source for base map information. An inherent advantage of digital data from government agencies is its cost. It is typically inexpensive. However, this is often offset by the data's accuracy and quality. Thematic coverages are often not up to date. However, it is important to note that specific characteristics of government data varies greatly across North America.

Attribute data has an even wider variety of data sources. Any textual or tabular data than can be referenced to a geographic feature, e.g. a point, line, or area, can be input into a GIS. Attribute data is usually input by manual keying or via a bulk loading utility of the DBMS software. ASCII format is a de facto standard for the transfer and conversion of attribute information.

The following figure describes the basic data types that are used and created by a GIS.




The basic data types that are used and created by a GIS (after Berry).

## Data Input Techniques

Since the input of attribute data is usually quite simple, the discussion of data input techniques will be limited to spatial data only. There is no single method of entering the spatial data into a GIS. Rather, there are several, mutually compatible methods that can be used singly or in combination.

The choice of data input method is governed largely by the application, the available budget, and the type and the complexity of data being input.

There are at least four basic procedures for inputting spatial data into a GIS. These are:

- Manual digitizing;
- Automatic scanning;
- Entry of coordinates using coordinate geometry; and the
- Conversion of existing digital data.

<u>Digitizing</u>

While considerable work has been done with newer technologies, the overwhelming majority of GIS spatial data entry is done by manual digitizing. A digitizer is an electronic device consisting of a table upon which the map or drawing is placed. The user traces the spatial features with a hand-held magnetic pen, often called a mouse or cursor. While tracing the features the coordinates of selected points, e.g. vertices, are sent to the computer and stored. All points that are recorded are registered against positional control points, usually the map corners, that are keyed in by the user at the beginning of the digitizing session. The coordinates are recorded in a user defined coordinate system or map projection. Latitude and longitude and UTM is most often used. The ability to adjust or transform data during digitizing from one projection to another is a desirable function of the GIS software. Numerous functional techniques exist to aid the operator in the digitizing process.

Digitizing can be done in a point mode, where single points are recorded one at a time, or in a stream mode, where a point is collected on regular intervals of time or distance, measured by an X and Y movement, e.g. every 3 metres. Digitizing can also be done blindly or with a graphics terminal. Blind digitizing infers that the graphic result is not immediately viewable to the person digitizing. Most systems display the digitized linework as it is being digitized on an accompanying graphics terminal.

Most GIS's use a spaghetti mode of digitizing. This allows the user to simply digitize lines by indicating a start point and an end point. Data can be captured in point or stream mode. However, some systems do allow the user to capture the data in an arc/node topological data structure. The arc/node data structure requires that the digitizer identify nodes.

Data capture in an arc/node approach helps to build a topologic data structure immediately. This lessens the amount of post processing required to clean and build the topological definitions. However, most often digitizing with an arc/node approach does not negate the requirement for editing and cleaning of the digitized linework before a complete topological structure can be obtained.

The building of topology is primarily a post-digitizing process that is commonly executed in batch mode after data has been cleaned. To date, only a few commercial vector GIS software offerings have successfully exhibited the capability to build topology interactively while the user digitizes.

Manual digitizing has many advantages. These include:

1. Low capital cost, e.g. digitizing tables are cheap;
2. Low cost of labour;
3. Flexibility and adaptability to different data types and sources;
4. Easily taught in a short amount of time - an easily mastered skill
5. Generally the quality of data is high;
6. Digitizing devices are very reliable and most often offer a greater precision that the data warrants; and
7. Ability to easily register and update existing data.

For raster based GIS software data is still commonly digitized in a vector format and converted to a raster structure after the building of a clean topological structure. The procedure usually differs minimally from vector based software digitizing, other than some raster systems allow the user to define the resolution size of the grid-cell. Conversion to the raster structure may occur on-the-fly or afterwards as a separate conversion process.

Automatic Scanning

A variety of scanning devices exist for the automatic capture of spatial data. While several different technical approaches exist in scanning technology, all have the advantage of being able to capture spatial features from a map at a rapid rate of speed. However, as of yet, scanning has not proven to be a viable alternative for most GIS implementation. Scanners are generally expensive to acquire and operate. As well, most scanning devices have limitations with respect to the capture of selected features, e.g. text and symbol recognition. Experience has shown that most scanned data requires a substantial amount of manual editing to create a clean data layer. Given these basic constraints some other practical limitations of scanners should be identified. These include :

1. hard copy maps are often unable to be removed to where a scanning device is available, e.g. most companies or agencies cannot afford their own scanning device and therefore must send their maps to a private firm for scanning;
2. hard copy data may not be in a form that is viable for effective scanning, e.g. maps are of poor quality, or are in poor condition;
3. geographic features may be too few on a single map to make it practical, cost-justifiable, to scan;
4. often on busy maps a scanner may be unable to distinguish the features to be captured from the surrounding graphic information, e.g. dense contours with labels;
5. with raster scanning there it is difficult to read unique labels (text) for a geographic feature effectively; and
6. scanning is much more expensive than manual digitizing, considering all the cost/performance issues.

Consensus within the GIS community indicates that scanners work best when the information on a map is kept very clean, very simple, and uncluttered with graphic symbology.

The sheer cost of scanning usually eliminates the possibility of using scanning methods for data capture in most GIS implementations. Large data capture shops and government agencies are those most likely to be using scanning technology.

Currently, general consensus is that the quality of data captured from scanning devices is not substantial enough to justify the cost of using scanning technology. However, major breakthroughs are being made in the field, with scanning techniques and with capabilities to automatically clean and prepare scanned data for topological encoding. These include a variety of line following and text recognition techniques. Users should be aware that this technology has great potential in the years to come, particularly for larger GIS installations.

Coordinate Geometry

A third technique for the input of spatial data involves the calculation and entry of coordinates using coordinate geometry (COGO) procedures. This involves entering, from survey data, the explicit measurement of features from some known monument. This input technique is obviously very costly and labour intensive. In fact, it is rarely used for natural resource applications in GIS. This method is useful for creating very precise cartographic definitions of property, and accordingly is more appropriate for land records management at the cadastral or municipal scale.

Conversion of Existing Digital Data

A fourth technique that is becoming increasingly popular for data input is the conversion of existing digital data. A variety of spatial data, including digital maps, are openly available from a wide range of government and private sources. The most common digital data to be used in a GIS is data from CAD systems. A number of data conversion programs exist, mostly from GIS software vendors, to transform data from CAD formats to a raster or topological GIS data format. Several ad hoc standards for data exchange have been established in the market place. These are supplemented by a number of government distribution formats that have been developed. Given the wide variety of data formats that exist, most GIS vendors have developed and provide data exchange/conversion software to go from their format to those considered common in the market place.

Most GIS software vendors also provide an ASCII data exchange format specific to their product, and a programming subroutine library that will allow users to write their own data conversion routines to fulfil their own specific needs. As digital data becomes more readily available this capability becomes a necessity for any GIS. Data conversion from existing digital data is not a problem for most technical persons in the GIS field. However, for smaller GIS installations who have limited access to a GIS analyst this can be a major stumbling block in getting a GIS operational. Government agencies are usually a good source for technical information on data conversion requirements.

Some of the data formats common to the GIS marketplace are listed below. Please note that most formats are only utilized for graphic data. Attribute data is usually handled as ASCII text files. Vendor names are supplied where appropriate.

| | |
|---|---|
| IGDS - Interactive Graphics Design Software (Intergraph / Microstation) | This binary format is a standard in the turnkey CAD market and has become a de facto standard in Canada's mapping industry. It is a proprietary format, however most GIS software vendors provide DGN translators. |
| DLG - Digital Line Graph (US Geological Survey) | This ASCII format is used by the USGS as a distribution standard and consequently is well utilized in the United States. It is not used very much in Canada even though most software vendors provide two way conversion to DLG. |
| DXF - Drawing Exchange Format (Autocad) | This ASCII format is used primarily to convert to/from the Autocad drawing format and is a standard in the engineering discipline. Most GIS software vendors provide a DXF translator. |
| GENERATE - ARC/INFO Graphic Exchange Format | A generic ASCII format for spatial data used by the ARC/INFO software to accommodate generic spatial data. |
| EXPORT - ARC/INFO Export Format . | An exchange format that includes both graphic and attribute data. This format is intended for transferring ARC/INFO data from one hardware platform, or site, to another. It is also often used for archiving.<br><br>ARC/INFO data. This is not a published data format, however some GIS and desktop mapping vendors provide translators. |

| | |
|---|---|
| | EXPORT format can come in either uncompressed, partially compressed, or fully compressed format |

A wide variety of other vendor specific data formats exist within the mapping and GIS industry. In particular, most GIS software vendors have their own proprietary formats. However, almost all provide data conversion to/from the above formats. As well, most GIS software vendors will develop data conversion programs dependant on specific requests by customers. Potential purchasers of commercial GIS packages should determine and clearly identify their data conversion needs, prior to purchase, to the software vendor.