Chapter 3

Data processing systems

3.1 Hardware and software trends

3.2 Geographic information systems

The handling of spatial data usually involves processes of data acquisition, storage and maintenance, analysis and output. For many years, this has been done using analogue data sources, manual processing and the production of paper maps. The introduction of modern technologies has led to an increased use of computers and digital information in all aspects of spatial data handling. The software technology used in this domain is geographic information systems.

3.2.1 The context of GIS usage

Spatial data handling involves many disciplines. We can distinguish disciplines that develop spatial concepts, provide means for capturing and processing of spatial data, provide a formal and theoretical foundation, are application-oriented, and support spatial data handling in legal and management aspects. Table 3.1 shows a classification of some of these disciplines. They are grouped according to how they deal with spatial information. The list is not meant to be exhaustive.

The discipline that deals with all aspects of spatial data handling is called geoinformatics. It is defined as:

> Geoinformatics is the integration of different disciplines dealing with spatial information.

3.2.2 GIS software

The main characteristics of a GIS software package are its analytical functions that provide means for deriving new geoinformation from existing spatial and attribute data. A GIS can be defined as follows [4]:

> A GIS is a computer-based system that provides the following four sets of capabilities to handle georeferenced data:
>
> 1. input,
> 2. data management (data storage and retrieval),
> 3. manipulation and analysis, and
> 4. output.

3.2.3 Software architecture and functionality of a GIS

According to the definition, a GIS always consists of modules for input, storage, analysis, display and output of spatial data. Figure 3.1 shows a diagram of these modules with arrows indicating the data flow in the system. For a particular GIS, each of these modules may provide many or only few functions. However, if one of these functions would be completely missing, the system should not be called a geographic information system.

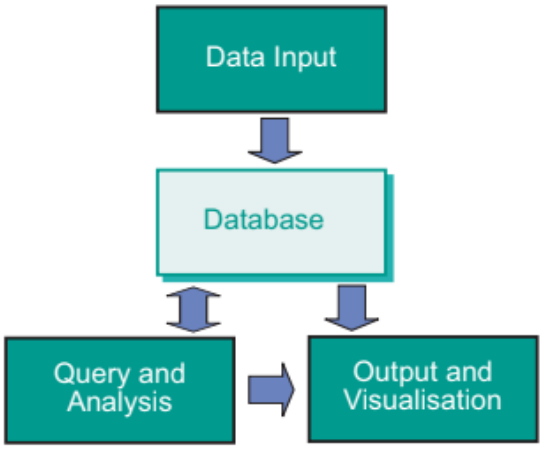

Figure 3.1: Functional components of a GIS

Method	*Devices*
Manual digitizing	• coordinate entry via keyboard • digitizing tablet with cursor • mouse cursor on the computer monitor (heads-up digitizing) • (digital) photogrammetry
Automatic digitizing	• scanner
Semi-automatic digitizing	• line following devices
Input of available digital data	• magnetic tape or CD-ROM • via computer network

Table 3.2: Spatial data input methods and devices used

Data input

The functions for data input are closely related to the disciplines of surveying engineering, photogrammetry, remote sensing, and the processes of digitizing, i.e., the conversion of analogue data into digital representations. Remote sensing, in particular, is the field that provides photographs and images as the raw base data from which to obtain spatial data sets. Additional techniques for obtaining spatial data are *manual digitizing*, *scanning* and sometimes *semi-automatic line following*.

Method	*Devices*
Hard copy	• printer • plotter (pen plotter, ink-jet printer, thermal transfer printer, electrostatic plotter) • film writer
Soft copy	• computer screen (CRT)
Output of digital data sets	• magnetic tape • CD-ROM • computer networks

Table 3.3: Data output and visualization

Data output and visualization

Data output is closely related to the disciplines of cartography, printing and publishing. Table 3.3 lists different methods and devices used for the output of spatial data.

3.2.4 Querying, maintenance and spatial analysis

The most distinguishing part of a GIS are its functions for spatial analysis, i.e., operators that use spatial data to derive new geoinformation.

Spatial queries and process models play an important role in satisfying user needs.

The combination of a database, GIS software, rules, and a reasoning mechanism (implemented as a so-called inference engine) leads to what is sometimes called a spatial decision support system (SDSS).

In a GIS, data are stored in layers (or themes).
Usually, several themes are part of a project.
The analysis functions of a GIS use the spatial and non-spatial attributes of the data in a spatial database to answer questions about the real world.

The following three classes are the most important query and analysis functions of a GIS, after [4]:

- Maintenance and analysis of spatial data,
- Maintenance and analysis of attribute data, and
- Integrated analysis of spatial and attribute data.

The first and third are GIS-specific, so are dealt with here. the second class is discussed in Section 3.3.

Questions	*Answers*	*GIS functions*
What is . . . ?	Display of data as maps, reports and tables, e.g., "What are the name and the address of the owner of that land parcel?"	Storage and query functions
What pattern . . . ?	Patterns in the data, e.g., all parcels with an area size greater than 2000.	Query functions with constraints
What . . . if . . . ?	A prediction about the data at a certain time or at a certain location.	Modelling functions

Table 3.5: Types of queries

Maintenance and analysis of spatial data

Maintenance of (spatial) data can best be defined as the combined activities to keep the data set up-to-date and as supportive as possible to the user community.

It deals with obtaining new data, and entering them into the system, possibly replacing outdated data. The purpose is have available an up-to-date, stored data set. After a major earthquake, for instance, we may have to update our digital elevation model to reflect the current elevations better so as to improve our hazard analysis.

Operators of this kind operate on the spatial properties of GIS data, and provide a user with functions as described below.

Format transformation functions convert between data formats of different systems or representations, e.g., reading a DXF file into a GIS.

Geometric transformations help to obtain data from an original hard copy source through digitizing the correct world geometry.
These operators transform device coordinates (coordinates from digitizing tablets or screen coordinates) into world coordinates (geographic coordinates, metres, etc.).

Map projections provide means to map geographic coordinates onto a flat surface (for map production), and vice versa.

Edge matching is the process of joining two or more map sheets. At the map sheet edges, feature representations have to be matched so as to be combined.

Graphic element editing allows to change digitized features so as to correct errors, and to prepare a clean data set for topology building.

Coordinate thinning is a process that often is applied to remove redundant vertices from line representations.

Integrated analysis of spatial and attribute data

Analysis of (spatial) data can be defined as computing from the existing, stored data set new information that provides insights we possibly did not have before. It really depends on the application requirements, and the examples are manifold. Road construction in mountainous areas is a complex engineering task with many cost factors such as the amount of tunnels and bridges to be constructed, the total length of the tarmac, and the volume of rock and soil to be moved. GIS can help to compute such costs on the basis of an up-to-date digital elevation model and soil map.

Functions of this kind operate on both spatial and non-spatial attributes of data, and can be grouped into the following types.

Retrieval, classification, and measurement functions

• Retrieval functions allow the selective search and manipulation of data without the need to create new entities.

• Classification allows assigning features to a class on the basis of attribute values or attribute ranges (definition of data patterns).

• Generalization is a function that joins different classes of objects with common characteristics to a higher level (generalized) class.

• Measurement functions allow measuring distances, lengths, or areas.

Overlay functions

belong to the most frequently used functions in a GIS application. They allow to combine two spatial data layers by applying the set theoretic operations of intersection, union, difference, and complement using sets of positions (geometric attribute values) as their arguments.
Thus we can find
• the potato fields on clay soils (intersection),
• the fields where potato or maize is the crop (union),
• the potato fields not on clay soils (difference),
• the fields that do not have potato as crop (complement).

Neighbourhood functions
operate on the neighbouring features of a given feature or set of features.

• Search functions allow the retrieval of features that fall within a given search window (which may be a rectangle, circle, or polygon).
• Line-in-polygon and point-in-polygon functions determine whether a given linear or point feature is located within a given polygon, or they report the polygons that a given point or line are contained in.

• The best known example of proximity functions is the buffer zone generation (or buffering). This function determines a fixed-width (or variable-width) environment surrounding a given feature.

• Topographic functions compute the slope or aspect from a given digital representation of the terrain (digital terrain model or DTM).

• Interpolation functions predict unknown values using the known values at nearby locations.

• Contour generation functions calculate contours as a set of lines that connect points with the same attribute value.
Examples are points with the same elevation (contours), same depth (bathymetric contours), same barometric pressure (isobars), or same temperature (isothermal lines).

Connectivity functions accumulate values as they traverse over a feature or over a set of features.
• Contiguity measures evaluate characteristics of spatial units that are contiguous (are connected with unbroken adjacency. Think of the search for a contiguous area of forest of certain size and shape.
• Network analysis is used to compute the shortest path (in terms of distance or travel time) between two points in a network (routing).
Alternatively, it finds all points that can be reached within a given distance or duration from a centre (allocation).
• Visibility functions are used to compute the points that are visible from a given location (viewshed modelling or viewshed mapping) using a digital terrain model.

3.3 Database management systems

A database management system (DBMS) is a software package that allows the user to set up, use and maintain a database. Like A GIS allows to set up a GIS application, a DBMS offers generic functionality for database organization and data handling.

3.3.1 Using a DBMS

There are various reasons why one would want to use a DBMS to support data storage and processing.

• **A DBMS supports the storage and manipulation of very large data sets.**
Some data sets are so big that storing them in text files or spreadsheet files becomes too awkward for use in practice. The result may be that finding simple facts takes minutes, and performing simple calculations perhaps even hours.

• **A DBMS can be instructed to guard over some levels of data correctness.**
For instance, an important aspect of data correctness is data entry checking: making sure that the data that is entered into the database is sensible data that does not contain obvious errors. Since we know in what study area we work, we know the range of possible geographic coordinates, so we can make the DBMS

check them.
The above is a simple example of the type of rules, generally known as integrity constraints, that can be defined in and automatically checked by a DBMS.
More complex integrity constraints are certainly possible, and their definition is part of the development of a database.

- **A DBMS supports the concurrent use of the same data set by many users.**
  Moreover, for different users of the database, different views of the data can be defined. In this way, users will be under the impression that they operate on their personal database, and not on one shared by many people. This DBMS function is called concurrency control.

- **A DBMS provides a high-level, declarative query language.**

  The most important use of the language is the definition of queries. A query is a computer program that extracts data from the database that meet the conditions indicated in the query. We provide a few examples below.

- **A DBMS supports the use of a data model.**
  A data model is a language with which one can define a database structure and manipulate the data stored in it. The most prominent data model is the relational data model. Its primitives are tuples (also known as records, or rows) with attribute values, and relations, being sets of similarly formed tuples.

- **A DBMS includes data backup and recovery functions to ensure data availability at all times.**
  As potentially many users rely on the availability of the data, the data must be safeguarded against possible calamities. Regular back-ups of the data set, and automatic recovery schemes provide an insurance against loss of data.
- **A DBMS allows to control data redundancy.**
  A well-designed database takes care of storing single facts only once. Storing a fact multiple times—a phenomenon known as data redundancy—easily leads to situations in which stored facts start to contradict each other, causing reduced usefulness of the data.

3.3.2 Alternatives for data management

A good question at this point is whether there are any alternatives to using a DBMS, when one has a data set to care about. Obviously, it all depends on how much data there is or will be, what type of use we want to make of it, and how many people will be involved.
On the small-scale side of the spectrum—when the data set is small, its use relatively simple, and with just one user—we might use simple text files, and a text processor. Think of a personal address book as an example, or a not-too-big batch of simple field observations.
If our data set is still small and numeric by nature, and we have a single type of use in mind, perhaps a spreadsheet program will do the job. This can be the case if we have a number of field observations with measurements that we want

to prepare for statistical analysis. However, if we carry out region- or nation-wide censuses, with many observation stations and/or field observers and all sorts of different measurements, one quickly needs a database to keep track of all the data. Spreadsheets also do not accommodate multiple uses of the same data set well.

All too often, we find that data collections—if they are made digital—reside in text files or spreadsheets, when the type(s) of use that the owner has in mind really requires a DBMS. Text files offer no support for data analysis whatsoever, except perhaps alphabetical ordering. Spreadsheets do support some data analysis, especially when it comes to calculations over a single table, like averages, sums, minimum and maximum values. All of such computations are, however, restricted to just a single table of data. When one wants to relate the values in the table with values of another nature in some other table, an expert hand and an effort in time are usually needed. It is precisely here where the knowledge of a good database query language pays off.