Unit Five

Data Source and Data Capturing

Unit objectives

At the end of this unit, you will be able to:

- ✓ Identify different types of data sources of GIS
- ✓ Identify methods of raster and vector data capture
- ✓ Identify methods of attribute data sources
- ✓ Understand methods of spatial and attribute data entry
- ✓ Understand source of GIS data errors
- ✓ Differentiate spatial data preparation methods.

5.1. Introduction

The processes of data collection are variously referred to as data entry, data acquisition, data capture, data conversion, data transfer, data translation, and digitizing. Although there are subtle differences between these terms, they describe the same thing, i**.e., adding geographic data into a database for analysis and map production.** It covers all aspects of transforming data captured from field observations, existing maps, or sensors into a GIS compatible digital format. This process of data encoding often referred to as **'data stream'**. It is a process of progressing from raw (analogue) data to an integrated digital GIS data. This unit describes data sources, data capturing methods, data encoding methods, error sources, some GIS data preparation methods.

5.2. Spatial data sources

Spatial data can be obtained from various sources. It can be collected from scratch, using direct acquisition techniques (as a primary data source), or indirectly by making use of the existing spatial data sources (as a secondary data source). *Primary data sources* are those data collected in digital format by direct measurement. However, it is not always feasible to obtain primary spatial data due to cost and available time. Hence, previously acquired (secondary data) may fit the current GIS need. *Secondary data sources* are digital and/or analog datasets that were originally captured for another purpose; and they need to be converted into a suitable digital format for the current GIS use. Table 5-1 shows some example primary and secondary spatial data sources.

Table 5-1: Classification of data sources

Source	Raster	Vector
Primary	➢ Digital remote sensing images ➢ Digital *satellite* images and aerial photographs	➢ GPS field measurement ➢ Survey field measurement
Secondary	➢ Scanned images or photographs ➢ Digital elevation models (DEM) generated from topographic contour maps	➢ Topographic maps

5.3. Methods of raster data capture

Raster data can be captured from different sources that may include analogue images, such as printed-paper maps or digital images available; usually captured using the following methods.

5.3.1. Satellite remote sensing (RS)

Remote sensing is the most popular form of primary raster data capture method. Data is collected using sensors that include cameras, digital scanners and LIDAR (Light Detection And Ranging), and platforms that consist of aircraft and satellites. Information derived from measurements of the amount of electromagnetic radiation reflected, emitted, or scattered. Typical products of RS data are orthophoto maps (geometrically corrected maps), satellite image maps, topographic maps, and thematic maps such as land-use and land-use change maps.

5.3.2. Aerial photography

Aerial photographs provided geospatial data for a wide range of applications since the early 20th century. There are two categories of aerial photographs: **vertical** and **oblique** photographs. **A vertical aerial photograph** is produced with a camera mounted into the floor of an aircraft. The resulting image is similar to a map and has a scale that is roughly constant throughout the image area. Vertical aerial photographs are more accurate for 2D mapping. **Oblique photographs** are obtained if the axis of the camera is not vertical in order to provide stereo imagery from overlapping pairs of images. Oblique images provide side views of objects such as buildings. Therefore, they are convenient for 3D mapping and image interpretation. Aerial photography needs ground control points visible on the photographs for georeferencing. This technique is called *photogrammetry*, which makes measurements from photographs to obtain the exact positions of surface points. Digital cameras and modern software ERDAS perform photogrammetry.

5.3.3. Satellite imagery

On-board satellite cameras continuously scan and take pictures of the earth. Satellite imagery data in conjunction with vector data, such as road networks, are widely used in GIS. NASA's Landsat7, Digital Globe's Quickbird and WorldView-1, GeoEye's IKONOS and OrbView, and the French SPOT satellite systems provide enormous amounts of high-resolution satellite imagery data in various seasons. An important feature of satellite and aerial photography systems is that they provide stereo imagery from overlapping pairs of images and used to create a 3D analogue or digital model from which 3D coordinates, contours, and DEM can be created.

5.3.4. Raster data capture using scanners

Scanning is the process of converting an analogue hardcopy map or image into a digital map or image. Analogue map is normally a printed-paper map; where as a digital map is stored in a known raster data format. This data can be transferred in to a computer and directly used as a raster backdrop, or a scanned map to be vectorized. Most GIS scanning require a resolution of 400–1000dpi (16-40dots/mm). Nevertheless, scanning does not result in structured datasets. After scanning, various image processing may apply to improve the scanned map quality. Table 5-2 shows map-scanning precautions. Scanners range from a small (A4) desktop scanner with a resolution of 200-800 dpi to high-end drum scanners for accurate large-sized (A0) documents.

Table 5-2: Map-scanning precautions for GIS use

Output quality	The scanned map needs to be sharp and clear. This can be improved (enhanced) by setting up of the brightness and contrast levels of an image.
Resolution	Usually measured in dpi. 150dpi for text, 300dpi for line maps and higher dpi for high quality orthophoto images.
Accuracy	Cleaning of stains/folding marks of a scanned map is essential before GIS use.
Georeferencing	If the scanned image is in low quality, distortion can create problem.
Vectorization	The resolution of the scanned map is very important to generate vector data.

5.4. Methods of vector data capture

Primary vector data capture is a major source of geographic data. The two main primary vector data capturing methods are surveying (ground) and GPS. Secondary vector data capture involves digitizing of vector objects from images, scanned maps, and other sources.

5.4.1. Surveying

Survey data sources could be analogue or digital surveying data. **Analogue survey data** are data derived from survey notes, and from observations of angles and distances of a survey campaign documents. **Digital survey data** are those data derived from field measurements of distances and angles stored digitally in a surveying instrument such as a total station; and then transferred as a digital file into GIS. Field surveys (also called ground surveys) usually employed for detail and accurate measurement of objects such as control stations, property boundaries, and building footprints. Field surveys include measurement and recording of distances, angles, and directions (positions) as well as preparation of sketches about areas and features being surveyed. Field surveying instruments, such as total station, measures angles and distances. It combines a digital theodolite, an electronic distance meter (EDM) and software running on an internal (or external) computer known as data collector. With the aid of triangulation, measured angles and distances used to calculate the coordinate positions of surveyed points. Other field survey instruments are theodolite measures horizontal and vertical angles, and laser range measures accurate distances.

5.4.2. Global Position System (GPS)

GPS is a satellite based surveying method provides an accurate geodetic coordinates of any point on the Earth at any time to a centimeter level of accuracy within few seconds. With specialized instruments, such as Real-Time-Kinematic (RTK) GPS, coordinates and heights can be determined and ported directly into the GIS software. GPS data provides location information and travel paths. New surveying systems perfectly combine a total station and GPS, called Smart Station. The entire software controlled via a keyboard. All measurements stored in the same database and displayed on screen.

5.4.3. Coordinate Geometry (COGO)

COGO data are precisely measured, often regarded as the only legally acceptable definition of land parcels, employing recording of bearings and distances. Source data may be the legal descriptions of parcels, records of survey notes, or parcel tract map documents. The data often used to supplement and update the existing data and for verification of data collected from aerial surveys and satellite remote sensing. Although COGO data obtained as part of a primary data capture, it is still a secondary source captured from hard copy map documents and survey notes.

5.5. Obtaining data from external sources - data exchange

The other GIS data source is data exchange from external sources. These days, GIS data is being shared among GIS users than ever before. New technologies increasing make the availability of geospatial data, particularly the Internet and stored in various databases (e.g. CSAs and NASA). Some data are freely available, and others are commercial, as in the case of most satellite imagery.

5.6. Methods of attribute data capture

Sources of attribute data include analogue text or tables such as printed or handwritten text or tables or fieldwork notes and digital text or tables available in a known format to be imported into the system. Soft ideas (e.g. expert knowledge) can be employed as attribute data source. Metadata is a special type of non-geometric data increasingly collected. GIS software derives some metadata automatically such as length and area of features.

5.7. Entering data into the GIS database

5.7.1. Survey and/or COGO data entry

Survey data entry includes keyboard entry; electronic data transfer/importing existing data or data from instruments; and data transfer from the GIS database. Data transfer from instruments is managed by using a cable device connected to computer. Existing survey data, which could be available in Excel or text formats, can be transferred in to a GIS database. COGO data can be entered using keyboard by converting bearings and distances into X, Y coordinates; or using Parcel Fabric Data Editor parcel traverse and COGO Tools.

5.7.2. Digitizing of vector data from scanned images

Digitizing is a process of converting paper map features into digital format. There are two types of digitizing techniques: manual digitizing and semi-automatic or automatic digitizing.

5.7.2.1. Manual digitizing

Manual digitizing is carried out in two modes. One is **point mode** where the operator digitizes a series of precise points or vertices of map features; and records them manually by changing directions of lines accurately. ArcMap connects vertices to create digital features. The other is **stream mode** where an ArcMap automatically adds vertices at intervals of time or distance as the operator moves around on the map. This mode is used to create curved lines, such as a river. But it is not as accurate as point mode. Stream mode requires more skill and it is faster than point mode digitizing. There are two methods of manual digitizing: **on-tablet** and **on-screen.**

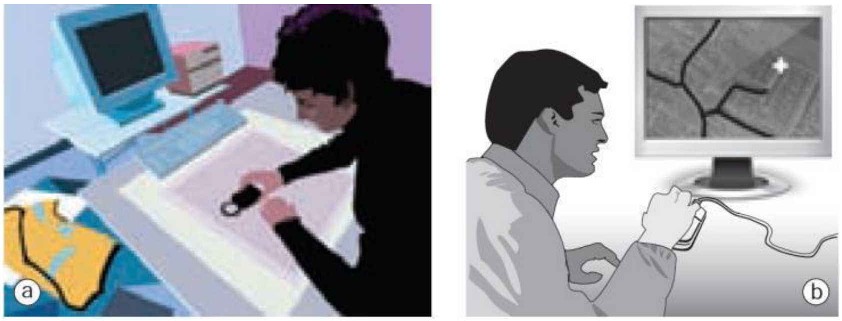

Figure 5-1: On-tablet digitizing (a); and on-screen digitizing (b)

With **on-screen manual digitizing** method, a scanned image or map is shown on a computer screen and digitized using a mouse or a pen. The method also called *'heads-up digitizing'* since a map is vertical and viewed without bending head down while digitizing. **On-tablet digitizing**, also called *'heads-down digitizing',* requires a digitizing tablet to fit a paper map on it. The operator uses a computer mouse-like device called pucks to follow lines on the map and enter into GIS. In both methods, the operator traces map features, and store location coordinates relative to known control points of the map. At least three control points are needed to *'lock'* the coordinate system onto the digitized data and to checking for any positional errors.

Nowadays, on-screen digitizing has replaced on-tablet digitizing due to the following reasons:

- More comfortable for the operator since the operator can see the digitizing process
- Obtain better positional accuracy by zooming in on the raster map on the screen.
- It is easy to keep track of finished vector objects and quality control.
- Its functionality is available in most standard GIS software.
- Possible to combine semi-automatic/automatic digitizing and editing techniques at a time

Manual digitizing requires a resolution of 200-300 dpi for printed maps, depending on the thinnest lines thickness; and a resolution of at least 800 dpi for aerial photographs. However, it is time-consuming (e.g. to digitize a complex contour map, it might take 100 person-hours); but its accuracy can never be better than the original hardcopy map. This is because the human hand positional accuracy level is about 40dpi at best and will decrease, as the operator gets bored.

5.7.2.2. Semiautomatic or automatic digitizing

Another set of techniques also works from a scanned image of the original map, but uses a GIS to find features in an image. They are known as semiautomatic or automatic digitizing, depending on how much operator interaction is required. If vector data is extracted from this procedure, a process known as vectorization. Vectorization process converts raster graphics to vector data using line-following software. Semiautomatic vectorization requires an operator to place a pointer at the start of a line, and the system automatically performs line following on the scanned map. Whenever there is a junction or intersection of lines on a map, it does not know how to continue. The operator has to *indicate the direction of lines to be digitized.* In automatic vectorization, the system recognizes map features and performs vectorization without operator's input. Vectorization process also called *skeletonizing*, involving three steps: (a) Line thinning—a process of reducing the quantity of line coordinate data; (b) Line extraction—a process of removing unnecessary duplicated lines; and (c) Topological reconstruction—construction of relationships.

ArcGIS has an extension called ArcScan that traces map lines automatically and creates vector data from map data. Vectorization requires a resolution of several pixels wide. For printed maps, a resolution of 300-600 dpi is sufficient. If the image is not clean, all imperfections on the map will be converted to false digital points, lines, or polygons. In this case, it is better to use manual digitizing. However, this method is less labor-intensive and much quicker than manual digitizing methods; but it can only be applied for relatively simple sources. The choice of a digitizing technique depends on the quality, complexity and contents of input document. Complex images are better digitized manually; simple images are better digitized automatically. Images that are full of detail and symbols, such as topographic maps and aerial photographs, are better digitized manually. Images that show only one type of information (e.g. contours) are better digitized automatically. However, in practice, the optimal choice is combination of methods. For example, contour line be automatically digitized, and then used to produce DEM.

5.7.3. Attribute data entry

Attribute data can be recorded with direct keyboard entry into a spreadsheet or database. Optical character recognition (OCR) or voice recognition also used for automatic entry of attribute data. Metadata is also entered in a similar way as other attribute data entered (Fazal, 2008). The essential requirement is a common identifier (or key) to relate object geometry and attributes together. This is important particularly when joining the two data types in a GIS database.

5.8. Sources of data errors

Spatial data errors are often more difficult to identify and correct. Errors may take many forms, depending on the data type, data processing steps, and method of data capture. Certain error types help to identify other errors. Table 5-3 shows example of vector data error sources.

Table 5-3: Examples of vector data error sources

Error	Description
Missing entities	Missing points, lines or boundary segments
Duplicate entities	Points, lines or boundary segments that have been digitized twice
Dislocated entities	Points, lines or boundary segments digitized in the wrong place
Missing labels	Unidentified polygons
Duplicate labels	Two or more identification labels for the same polygon
Artefacts of digitizing	Undershoots, overshoots, wrongly placed nodes, loops and spikes

5.9. Data preparation

Spatial data preparation aims at making of acquired spatial data fit for the GIS. Corrections can be done interactively by the operator **'on-screen',** or **automatically** by the GIS software. The following are among others important GIS data preparation operations and functions:

- Data clipping; add missing boundary
- Changing formats and projections;
- Combine multiple polygons;
- Cleaning and checking attributes;
- Correcting gaps and overlaps;
- Removal of redundant vertices of line features (called coordinate thinning);
- Building feature topology;
- Calculate attribute feature attribute values such as areas and lengths.

Data may require also conversion between vector and raster formats to match with other datasets. Data preparation process also includes associating attribute data with the spatial features through either manual input or reading digital attribute files.

5.9.1. Data checks and repairs

Acquired datasets must be checked for their quality in terms of accuracy, consistency and completeness. Vector data may require editing, such as trimming of overshoot lines, deleting duplicate lines, closing gaps in lines, and generating polygons. Some GIS operations, such as line smoothing and data clean-up (removing duplicate lines), may identify and automatically correct errors.'Clean-up' operations often performed in a standard sequence. For example, split crossing lines **then** create nodes at intersections **then** erase dangling lines **then** generate polygons. With polygon data, sometimes, polylines do not connect to form closed boundaries. Therefore, they must be connected to develop topology of the polygons.

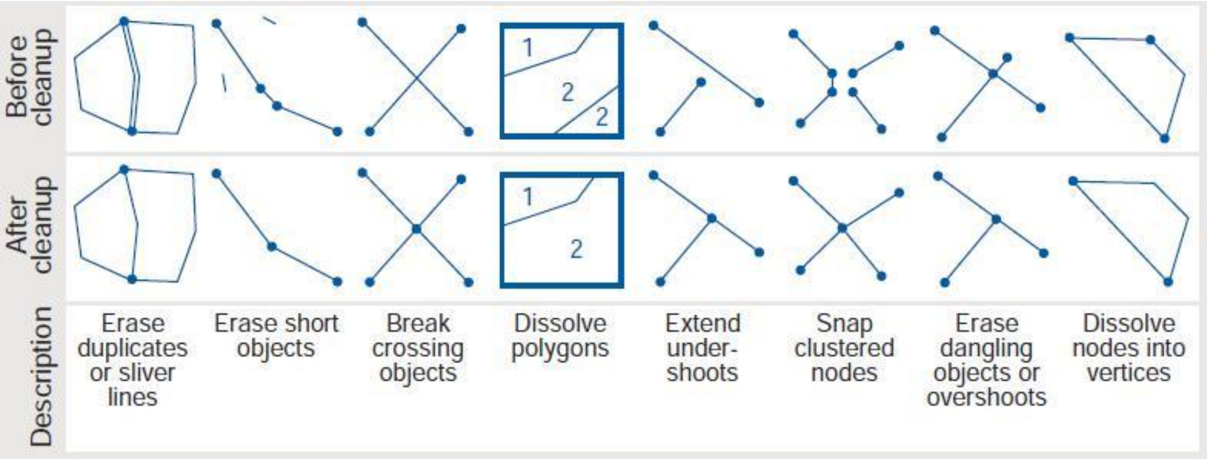

Figure 5-2: Clean-up operations for vector data

Rasterization and Vectorization	Vector data may be converted to raster data to carry out spatial data analysis on them. This process, called rasterization. To avoid information loss, the raster resolution should be carefully chosen. Vectorization may causes errors of small spikes along lines, rounded corners, errors in T- and X-junctions, displaced lines or jagged curves. These can be corrected in automatically or interactively.
Associating attributes	Vector data can automatically associate attributes to features with their unique identifiers. In raster, attributes are assigned to all cells.
Topology generation	Used to support data editing and ensure data quality. Typically, topology rules are first defined (e.g. ’there should be no gaps between polygons’), after rules validated. Errors are identified automatically, and corrected by manual editing.

5.9.2. Combining data from multiple sources

GIS usually involves multiple datasets. They may be of the same area but differ in accuracy, or they may be of adjacent areas but merged into a single dataset, or the datasets may be of the same or adjacent areas but referenced in different coordinate systems. All need to be corrected.

Differences in accuracy: - Images may come at a certain resolution, and printed maps at a certain scale. This typically results in differences of resolution of acquired datasets. If two polygons of digitized maps are at different scales overlaid, they do not perfectly coincide, and polygon boundaries cross each other. This causes small, artefact polygons in the overlay known as sliver polygons. If the map scales involved differ significantly, the polygon boundaries of the large-scale map should probably take priority, but when the differences are slight.

Merging datasets of adjacent areas: - separately digitized adjacent datasets need to be merged or combined into a single **'seamless'** or homogeneous dataset. GIS has **merge** or **edge matching** functions to solve such problems. *Edge matching* is a process of joining two or more map sheets. It resolves mismatches of boundaries; rebuilding of topology; and deleting or dissolving of redundant boundary lines. This may require visual check and interactive editing.

Differences in coordinate systems: - data layers to be combined or merged may be referenced to different coordinate systems or datums. Such data need coordinate transformation, or both coordinate and datum transformation. Scanned maps may require georeferencing and projection. Digital photogrammetry may requires orthorectification to produce orthophotos or vector maps.