Chapter 8 – GIS Data Collection

§ 1 Introduction

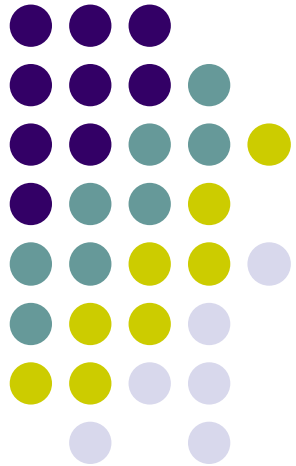
§ 2 Primary geographic data capture

§ 3 Secondary geographic data capture

§ 4 Obtaining data from external (data transfer)

§ 5 Capturing attribute data

§ 6 Managing a data collection project



Learning Objectives

- After reading this chapter you will be able to:
  - Describe data collection workflows
  - Understand the primary data capture techniques in remote sensing and surveying
  - Be familiar with the secondary data capture techniques of scanning, manual digitizing, vectorization
  - Understand the principles of data transfer, sources of digital geographic data, and geographic data formats
  - Analyze practical issues associated with managing data capture projects

§ 1 Introduction

- GIS can contain geographic data from many diverse sources by
  - data capture (direct data input)
  - data transfer (input of data from other systems)
- Primary sources are captured specifically by direct measurement
- Secondary sources are those reused from earlier studies or obtained from other systems
  - obtained in either digital or analog format
- Data capture, data automation, data conversion, data transfer, data translation, and digitizing essentially describe the same thing

Table 9.1 Classification of geographic data for data collection purposes with examples of each type

	Raster	Vector
Primary	Digital satellite remote-sensing images	GPS measurements
	Digital aerial photographs	Survey measurements
Secondary	Scanned maps or photographs	Topographic maps
	Digital elevation models from topographic map contours	Toponymy (placename) databases

§1 Introduction（cont.）

- Need to have tools to transform spatial data of various types into digital format
- Data capture is a major bottleneck in application of GIS technology
  - Costs of data capture often consume 85% or more of project costs
  - Data capture and input is labor intensive, tedious, error-prone
  - Essential to find ways to reduce costs, maximize accuracy
  - Sharing of digital data is one way around the bottleneck

Table 9.2 Breakdown of costs (in $1000s) for two typical client-server GIS as estimated by the authors

	10 seats		100 seats	
	$	%	$	%
Hardware	30	3.4	250	8.6
Software	25	2.8	200	6.9
Data	400	44.7	450	15.5
Staff	440	49.1	2000	69.0
Total	895	100	2900	100

§ 1 Introduction(cont.)

§ 1.1 Data collection workflow

- Data collection involves a series of stages
  - ***Planning*** includes establishing user requirements, garnering resources (staff, hardware, and software), and developing a project plan
  - ***Preparation*** involves many tasks such as obtaining data, redrafting poor-quality map sources, editing scanned map images, and removing noise. It may also involve setting up appropriate GIS hardware and software systems to accept data
  - ***Digitizing and transfer*** are the stages where the majority of the effort will be expended
  - ***Editing and improvement*** covers many techniques designed to validate data, as well as correct errors and improve quality
  - ***Evaluation*** is the process of identifying project successes and failures

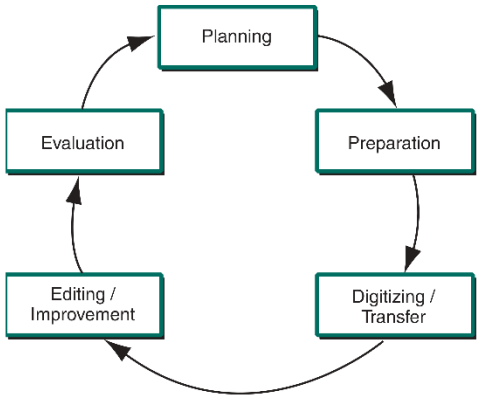
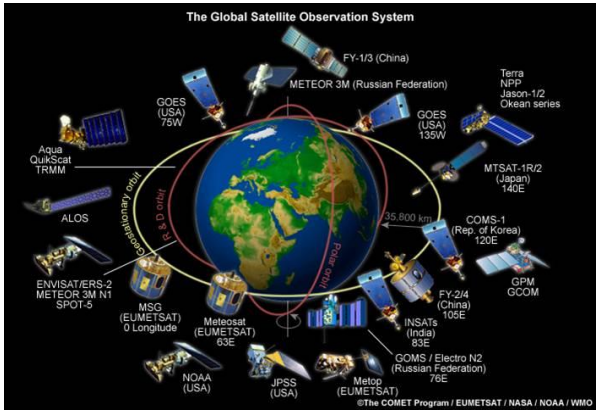


Figure 9.1 Stages in data collection projects

§ 2 Primary geographic data capture

- Remote sensing is the measurement of physical, chemical, and biological properties of objects without direct contact
  - Three key aspects of resolution are: spatial, spectral, and tempora
  - Other new RS technique: Radar data , Lidar, InSAR etc
  - Photogrammetry can capture higher resolution data such as urban 3D model
- The two main branches of vector data capture are ground surveying and GNSS
- Fieldwork by geographers or other scientists is the very important data source

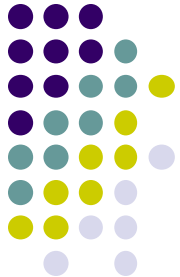


Remote sensing


Ground Surveying


Fieldwork

§ 3 Secondary geographici data capture



- Secondary sources is to create databases from maps, photographs, and other hard-copy documents
  - Most traditional spatial data are recorded on the maps or other analog media
  - Has gone through several decades to digitize maps or aerial photos into database
- Scanning is used to capture raster data. Table digitizing, heads-up digitizing, stereo-photogrammetry are used for vector data

§ 3 Secondary geographici data capture(cont.)

- Modes of data input:
  - Keyboard entry for non-spatial attributes and occasionally locational data
  - Manual locating devices
    - user directly manipulates a device whose location is recognized by the computer ,e.g. digitizing
  - Automated devices
    - automatically extract spatial data from maps and photography , e.g. scanning
  - Conversion directly from other digital sources

§ 3 Secondary geographici data capture(cont.)

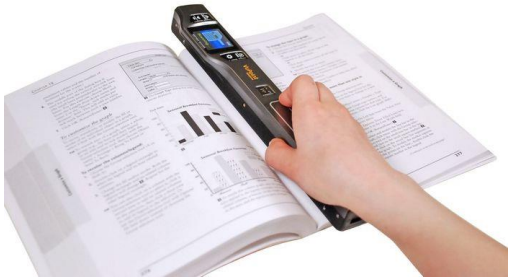
§ 3.1 Raster data capture using scanners

- Electromechanical systems are typically more expensive and slower, but can create better quality products
- one common class of scanners involves attaching the graphic to a drum
  - as the drum rotates about its axis, a scanner head containing a light source and photodetector reads the reflectivity of the target graphic, and digitizing this signal, creates a single column of pixels from the graphic
  - the scanner head moves along the axis of the drum to create the next column of pixels, and so on through the entire scan

Drum Scanner

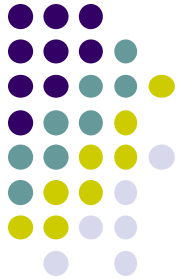
§ 3 Secondary geographici data capture(cont.)

- this controls distortion by bringing the single light source and detector to position on a regular grid of locations on the graphic
- systems may have a scan spot size of as little as 25 micrometers, and be able to scan graphics of the order of 1 meter on a side
- an alternative mechanism involves an array of photodetectors which extract data from several rows of the raster simultaneously
  - the detector moves across the document in a swath
  - when all the columns have been scanned, the detector moves to a new swath of rows



Flatbed scanner and handheld scanner

§ 3 Secondary geographici data capture(cont.)

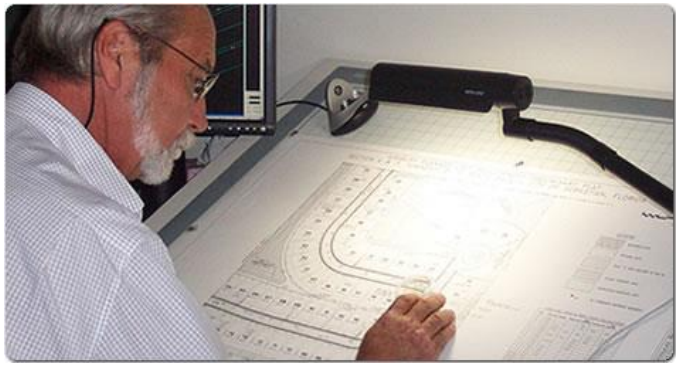


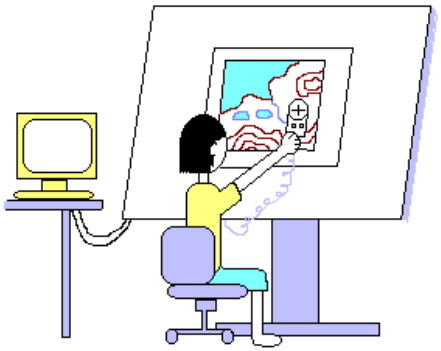
- Documents must be clean (no smudges or extra markings)
- Lines should be at least 0.1 mm wide
- Complex line work provides greater chance of error in scanning
- Text may be accidently scanned as line features
- Contour lines cannot be broken with text
- Automatic feature recognition is not easy (two contour lines vs. road symbols) diagram
- Special symbols (e.g. marsh) must be recognized and dealt with
- If good source documents are available, scanning can be an efficient time saving mode of data input

§ 3 Secondary geographici data capture(cont.)

§ 3.2 Vector data capture using manual digitizing

- The position of an indicator as it is moved over the surface of the digitizing tablet is detected by the computer and interpreted as pairs of x,y coordinates
- There are control buttons on the cursor which permit control of the system without having to turn attention from the digitizing tablet to a computer terminal

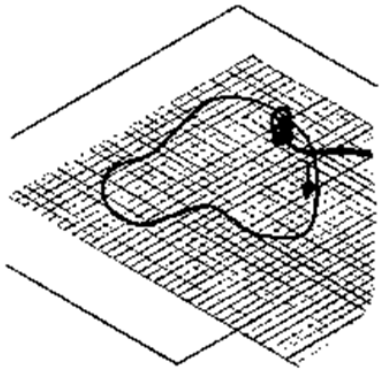




Digitizer and manual digitizing

§ 3 Secondary geographici data capture(cont.)

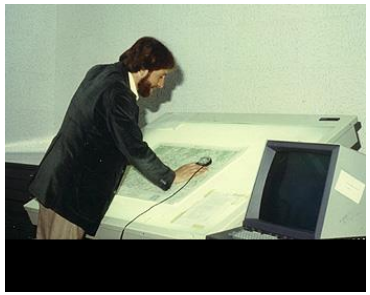
- Tablets use a grid of wires embedded in the tablet to generate a magnetic field which is detected by the cursor
  - accuracies are typically better than 0.1 mm
  - this is better than the accuracy with which the average operator can position the cursor
  - functions for transforming coordinates are sometimes built into the tablet and used to process data before it is sent to the host

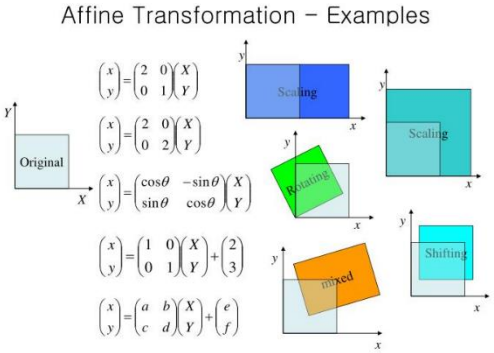


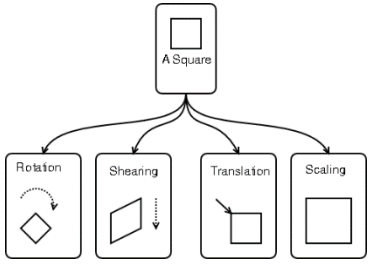
Digitizer tablet

§ 3 Secondary geographici data capture(cont.)

- The map is affixed to a digitizing table
- three or more control points ("reference points", "tics", etc.) are digitized for each map sheet
  - easily identified points such as four corners of a map
  - the coordinates of these points will be known in the final database
  - the control points are used by the system to calculate the mathematical transformationsto convert all coordinates to the final system
    - We can use affine transformation to establish function between theoretical points and measured points(at least 3 pairs of coordinates needed)
      - $x' = ax + by + p$
      - $y' = cx + dy + q$





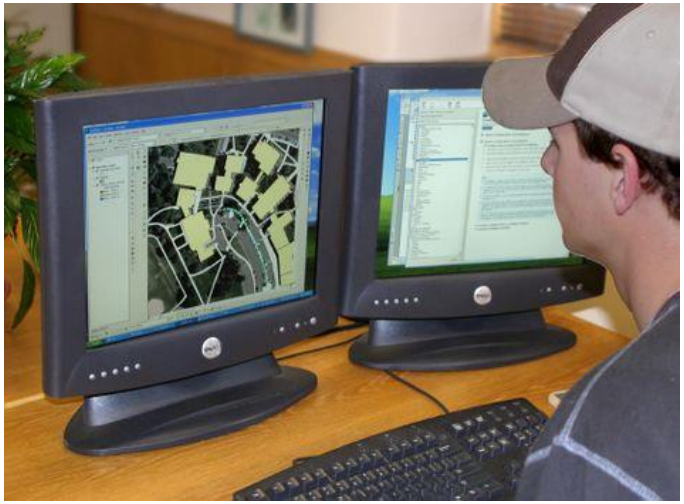


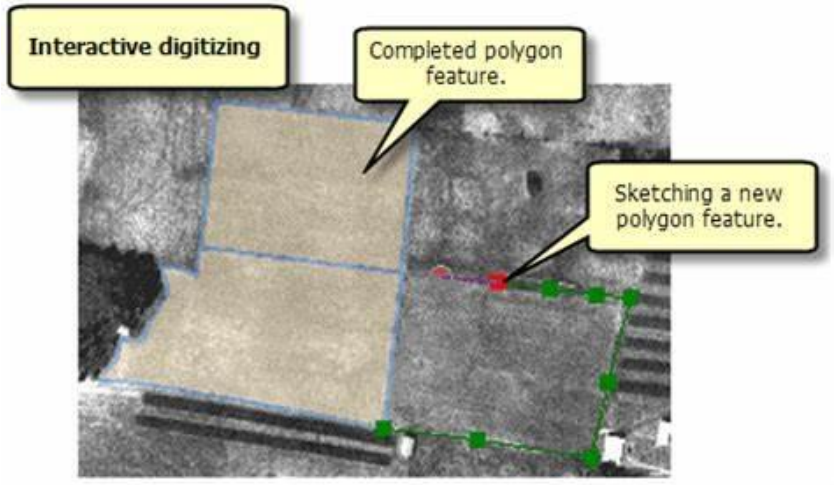
Digitizing operation and map affix

§ 3 Secondary geographici data capture(cont.)

§ 3.3 Heads-up digitizing and vectorization

- Vectorization is the process of converting rasterdata into vector data. The reverse is called rasterization
- Heads-up digitizing is a combination of scanning and manual digitizing by digitizing vector objects manually straight off a computer screen using a mouse or digitizing cursor





Heads-up digitizing

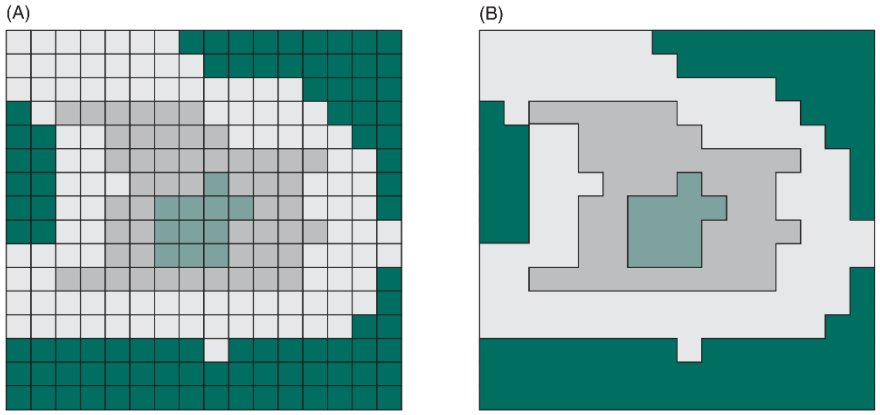
§ 3 Secondary geographici data capture(cont.)

- Can use graphic software such as CorelDraw to do it in three step:
  - Scan the map or aerial photo
  - Digitizing the map by zooming in or out
  - Rectify the digitized map with affine transformation

Heads-up digitizing

§ 3 Secondary geographici data capture(cont.)

- Use software to perform automated vectorization in either batch or semi-interactive mode is possible
- Batch vectorization takes an entire raster file and converts it to vector objects in a single operation
- Vector objects are created using software algorithms that build simple (spaghetti) line strings from the original pixel value
- Batch vectorization is best suited to simple bi-level maps of, for example, contours, streams, and highways



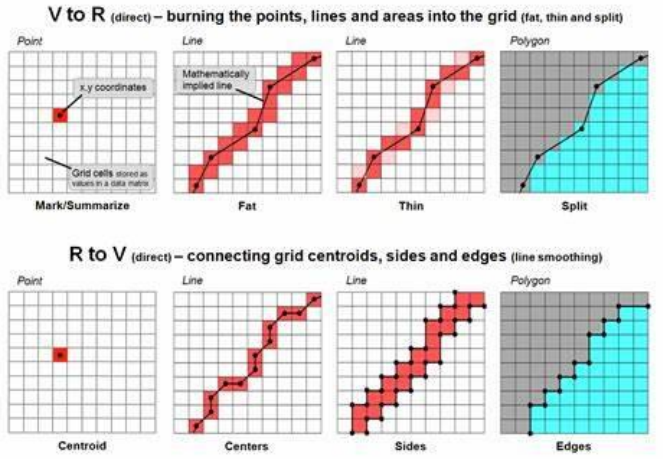


Figure 9.8 Batch vectorization of a scanned map: (A) original raster file; (B) vectorized polygons. Adjacent raster cells with the same attribute values are aggregated. Class boundaries are then created at the intersection between adjacent classes in the form of vector lines

§ 4 Data transfer

- The best way to find geographic data is to search the Internet using one of the specialist geolibraries or SDI geographic data geoportals
- Many website can provide dataset download or data service
- Data can be transferred between systems by direct read into memory or via an intermediate file format
  - AutoCAD DWG, DXF
  - Adobe AI, PS, PDF
  - ESRI Shapefile
  - Google KML, KMZ
- Data exchange can be very important for GIS
- Data exchange standard can solve it

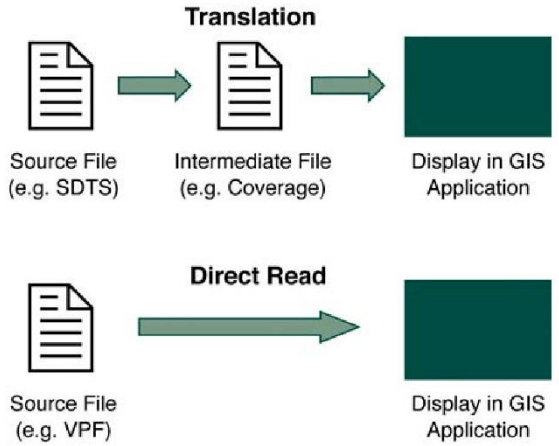


Figure 9.16 Comparison of data access by translation and direct read

§ 6 Managing a data collection project

- There is a fundamental tradeoff between quality, speed, and price
- Collecting high-quality data quickly is possible, but it is also very expensive. If price is a key consideration then lower-quality data can be collected over a longer period
- A key decision facing managers of such projects is whether to pursue a strategy of incremental or very rapid collection
- Whichever approach is preferred, a pilot project carried out on part of the study area and a selection of the data types can prove to be invaluable
- A further important decision is whether data collection should use in-house or external resources

§ 5 Capturing attribute data

- Attributes can be entered by direct data loggers, manual keyboard entry, optical character recognition (OCR) or, increasingly, voice recognition
- Much the most common method is direct keyboard data entry into a spreadsheet or database
- For some projects, a custom data entry form with in-built validation is preferred
- A common identifier (also called a key) that can be used to relate object geometry and attributes together
- Metadata(元数据） is data about data, collecting in the same way as attribute data