9 GIS Data Collection

OVERVIEW

This chapter reviews the main methods of GIS data capture and transfer and introduces key practical management issues.

It distinguishes between primary (direct measurement) and secondary (derivation from other sources) data capture for both raster and vector data types.

LEARNING OBJECTIVES

- **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, photogrammetry, and COGO feature construction;**
- **Understand the principles of data transfer, sources of digital geographic data, and geographic data formats;**
- **Analyze practical issues associated with managing data capture projects.**

KEY WORDS AND CONCEPTS

Data capture, data transfer, primary and secondary data sources, resolution (spatial, spectral and temporal), scanning, digitizing, error, photogrammetry, COGO, data transfer, data formats, ISO, CEN, OGC, OCR

OUTLINE

9.1 Introduction
9.2 Primary geographic data capture
9.3 Secondary geographic data capture
9.4 Obtaining data from external sources (data transfer)
9.5 Capturing attribute data
9.6 Citizen-centric Web based Data Collection
9.7 Managing a data collection project

CHAPTER SUMMARY

9.1 Introduction

In this chapter, data collection is split into *data capture* (direct data input) and *data transfer* (input of data from other systems).

- Two main types of data capture are
  - *Primary data sources* are those collected in digital format specifically for use in a GIS project.
  - *Secondary sources* are digital and analog datasets that were originally captured for another purpose and need to be converted into a suitable digital format for use in a GIS project.
- This chapter describes the data sources, techniques, and workflows involved in GIS data collection.
- The processes of data collection are also variously referred to as data capture, data automation, data conversion, data transfer, data translation, and digitizing.
- Table 9.2 shows a breakdown of costs for two typical client-server GIS implementations.
- Data collection is a time consuming, tedious, and expensive process.
- Typically it accounts for 15–50% of the total cost of a GIS project
- If staff costs are excluded from a GIS budget, then in cash expenditure terms data collection can be as much as 60–85% of costs.

9.1.1 Data collection workflow

- Figure 9.1 shows the stages in data collection projects
- *Planning* includes establishing user requirements, garnering resources, and developing a project plan.

- *Preparation* involves obtaining data, redrafting poor-quality map sources, editing scanned map images, removing noise, 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.

9.2 Primary geographic data capture

9.2.1 Raster data capture

- Remote sensing is a technique used to derive information about the physical, chemical, and biological properties of objects without direct physical contact
- Information is derived from measurements of the amount of electromagnetic radiation reflected, emitted, or scattered from objects.
- Figure 9.2 shows the spatial and temporal characteristics of commonly used remote sensing systems and their sensors
- *Resolution* is a key physical characteristic of remote sensing systems.
- *Spatial resolution* refers to the size of object that can be resolved and the most usual measure is the pixel size.
- *Spectral resolution* refers to the parts of the electromagnetic spectrum that are measured.
- *Temporal resolution*, or repeat cycle, describes the frequency with which images are collected for the same area.
- A paragraph describes SPOT imagery
- Aerial photography is equally important in medium- to large-scale projects
- Photographs are normally collected by analog optical cameras and later scanned
- Aerial Photographs are usually collected on an ad hoc basis
- Can provide stereo imagery for the extraction of digital elevation models
- Advantages are
    - Consistency of the data
    - Availability of systematic global coverage
    - Regular repeat cycles
- Disadvantages are
    - Resolution is often too coarse

- Many sensors are restricted by cloud cover

9.2.2 Vector data capture

- Two main branches are ground surveying and GPS
  - Distinction is increasing blurred

9.2.2.1 Surveying

- Ground surveying is based on the principle that the 3-D location of any point can be determined by measuring angles and distances from other known points.
- Traditional equipment like transits and theodolites have been replaced by total stations that can measure both angles and distances to an accuracy of 1 mm
- Ground survey is a very time-consuming and expensive activity, but it is still the best way to obtain highly accurate point locations.
- Typically used for capturing buildings, land and property boundaries, manholes, and other objects that need to be located accurately.
- Also employed to obtain reference marks for use in other data capture projects

9.2.2.2 LiDAR

- Relatively new technology that employs a scanning laser rangefinder to produce accurate topographic surveys
- Typically carried on a low-altitude aircraft that also has an inertial navigation system and a differential GPS to provide location.

9.3 Secondary geographic data capture

9.3.1 Raster data capture using scanners

Three main reasons to scan hardcopy media are

- Documents are scanned to reduce wear and tear, improve access, provide integrated database storage, and to index them geographically
- Film and paper maps, aerial photographs, and images are scanned and georeferenced so that they provide geographic context for other data
- Maps, aerial photographs and images are scanned prior to vectorization

9.3.2 Vector data capture

- Secondary vector data capture involves digitizing vector objects from maps and other geographic data sources.

9.3.2.1 Heads-up digitizing and vectorization

- Vectorization is the process of converting raster data into vector data.

- The simplest way to create vectors from raster layers is to digitize vector objects manually straight off a computer screen using a mouse or digitizing cursor.
- Describes how automated vectorization is performed

9.3.2.2 Measurement error

- Figure 9.10 presents some examples of human errors that are commonly introduced in the digitizing procedure including overshoots, undershoots, invalid polygons, and sliver polygons
- Discussion of how errors may arise by the use of rubbersheeting which assumes that spatial autocorrelation exists among errors

9.3.2.3 Photogrammetry

- Is the science and technology of making measurements from pictures, aerial photographs, and images.
- Measurements are captured from overlapping pairs of photographs using stereo plotters.
- Figure 9.13 shows a typical workflow in digital photogrammetry
- Orientation and triangulation are fundamental photogrammetry processing tasks.
    - Orientation is the process of creating a stereo model suitable for viewing and extracting 3-D vector coordinates that describe geographic objects.
    - Triangulation (also called ‘block adjustment’) is used to assemble a collection of images into a single model so that accurate and consistent information can be obtained from large areas.
- Orthoimages are images corrected for variations in terrain using a DEM.
- Photogrammetry is a very cost-effective data capture technique that is sometimes the only practical method of obtaining detailed topographic data

9.3.2.4 COGO data entry

- COGO is a contraction of the term *coordinate geometry*, a methodology for capturing and representing geographic data.
- COGO uses survey-style bearings and distances to define each part of an object
- COGO data are very precise measurements and are often regarded as the only legally acceptable definition of land parcels.

9.4 Obtaining data from external sources (data transfer)

A small selection of key data sources is listed in Table 9.3

The best way to find geographic data is to search the Internet

9.4.1 Geographic data formats

- One of the biggest problems with data obtained from external sources is that they can be encoded in many different formats.
- Many tools have been developed to move data between systems and to reuse data through open application programming interfaces (APIs).
- More than 25 organizations are involved in the standardization of various aspects of geographic data and geoprocessing
- ISO (International Standards Organization) is responsible for coordinating efforts through the work of technical committees TC 211 and 287
- In Europe, CEN (Comité Européen de Normalisation) is engaged in geographic standardization.
- OGC (Open Geospatial Consortium) is a group of vendors, academics, and users interested in the interoperability of geographic systems
- Geographic data translation software must address both syntactic and semantic translation issues.
- Syntactic translation involves converting specific digital symbols (letters and numbers) between systems.
- Semantic translation is concerned with converting the meaning inherent in geographic information.
- While the former is relatively simple to encode and decode, the latter is much more difficult and has seldom met with much success to date.

9.5 Capturing attribute data

- Attributes can be entered by direct data loggers, manual keyboard entry, optical character recognition (OCR) or, increasingly, voice recognition.
- An essential requirement for separate data entry is a common identifier (also called a key) that can be used to relate object geometry and attributes together following data capture

9.6 Citizen centric web based data collection

- Describes how a raft of new Web 2.0 technologies has enabled organizations and individual projects to use citizens to collect data across a wide variety of thematic and geographic areas

9.7 Managing a data collection project

- Most of the general principles for any GIS project apply to data collection: the need for a clearly articulated plan, adequate resources, appropriate funding, and sufficient time.
- A key decision facing managers of such projects is whether to pursue a strategy of incremental or very rapid collection.
- A further important decision is whether data collection should use in-house or external resources.

ESSAY TOPICS

1. Distinguish between primary and secondary data and give examples of each. In what circumstances is this distinction difficult to maintain?
2. Why is data maintenance often a far more difficult and expensive activity than the initial data collection?
3. What do you understand by the terms ‘active’ and ‘passive’ satellite sensor systems and what are the relative advantages of each?
4. Why is it often necessary to scan paper documents for data entry into a GIS?
5. Describe the necessary steps in a workflow for manual digitizing using a semi-automatic digitizer. How and why does this process introduce ‘error’ into the database?
6. You are required to merge together in your GIS database digital cartographic data with some satellite imagery. What are the necessary steps in this process and the likely sources of difficulty?
7. How does national and international legislation on freedom of information and copyright affect the market for geospatial data?
8. What are the difficulties in translating between different data formats, and what software solutions have been suggested?
9. It is often suggested that in satellite imagery there is a trade off between spatial, spectral and temporal resolution. Outline and illustrate what is meant by these properties. To what extent do the data in Table 9.2 support this idea?
10. Describe the various ways by which ‘error’, defined as the difference between reality and our representation of it, can be introduced in the process of data collection and integration into a GIS.