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.

MULTIPLE CHOICE QUESTIONS (MCQ)

1. The table below lists examples of geographic data used in GIS. For each example, state whether it is usually (a) raster or vector and (b) a primary or a secondary source and (c) digital or analogue. In some cases your answer can be 'both':

Geographical data	Raster or vector?	Primary or secondary?	Digital or analogue?
Satellite imagery			
Aerial photography			
Printed maps			
Elevation models			
GPS measurements			
Survey measurements			
Place name data			
Census bureau records			

2. Rank the following satellite platforms in increasing order (i.e. least frequent = 1) of temporal resolution of their imagery:

Platform	Answer
Landsat	
SPOT	
Meteosat	
IKONOS	

3. For each of the four listed elements of the cost of a client-server GIS to serve 100 users, match it to its estimated percentage of the total costs of the system:

Cost element	ANSWER a, b, c or d	Choose from:
Hardware		a) 15.5
Software		b) 69.0
Data		c) 6.9
Staff		d) 8.6

4. If we exclude staff costs from a GIS budget, what is the likely maximum cash expenditure on data acquisition? Is it a) 15% b) 30% or c) 85% of the total budget? ANSWER ….

5. In the data collection cycle, place the following items in their logical sequence:

Process	**Order (1) to (4)**
Editing and improvement	
Planning	
Digitizing/transfer	
Planning	
Evaluation	

6. Which of the following methods of sensing the environment are remote and which are direct?

Method	**Answer Remote or direct**
Aerial Photography	
Satellite imagery	
Standard thermometer	
GPS	
pH meter	

7. You have a scanner capable of recording using 8 binary digit 'bit's for each pixel in its output. Which of the following hardcopy media would it be capable of recording accurately:

Medium	**Answer YES or NO**
Newspaper printed gray scale	
Color aerial photograph	
CAD drawing	
USGS 1:24000 Quad sheet (Figure 3.14)	

8. Match each of the listed earth observation systems to the spatial resolution of its data:

System	Answer (a)-(f)	Choose from:
GEOS (Viz)		a) 0.25 x 0.25m
Digital aerial photography		b) 5 x 5km
Landsat 4 MSS		c) 80 x 80m
Meteosat		d) 1 x 1 km
Landsat 4 TM		e) 30 x 30m
SPOT Panchromatic		f) 10 x 10m

9. You need to acquire data to populate a GIS using a variety of secondary data providers. In the table below, match each data type to the most probable source:

Type	Answer (a)-(f)	Possible providers
Lifestyle classifications		a) National census agency
Population		b) Military/National mapping agency
Toponymy		c) National mapping agency
Elevation		d) Commercial and military
Satellite imagery		e) National government
Administrative areas		f) Private specialist agency

10. Which of the following is the most essential in linking attribute and geographic components in a GIS database?
    a. Complete records
    b. Metadata describing both
    c. A key field
    d. Lineage information

CLASS AND INDIVIDUAL ACTIVITIES

1. Use the web to discover what satellite imagery and digital aerial photography are available for a 20 x 20km area centered on your home or place of study. Make a table listing the spatial, spectral and temporal resolution of each source together with its ease of acquisition, georeferencing and other relevant metadata.

2. Table 9.4 lists a dozen data aggregation or indexing websites. Make a systematic visit to all to determine what data are available for your home area.
3. Choose either a high resolution aerial photograph or satellite image of a small area known to you and select a few polygonal objects that appear on it, such as woodland, field boundaries and the like. Use an appropriate tool to 'heads up' digitize the outline from the screen. Pixel co-ordinates can be obtained using a simple picture editing program such as Microsoft's Paint™. Make a list of the errors you have introduced into the data. Having obtained these picture values, consider how you would register them onto some known map projection. The next activity is a possible extension.
4. A critical and sometimes neglected step in many data integration exercises is the co-registration of the map and image data onto the same co-ordinate system. O'Sullivan and Unwin (2010, pages 221-324) provide more detail. Choose any suitable satellite image of an area for which you also have good fine scale mapping, and use the 'tick point' approach to co-register a series of points on the image onto the map co-ordinates. This exercise can be done using GIS software, but it is very instructive to follow all the steps using more basic tools such as a spreadsheet or statistical analysis program. It is probable that a simple affine transformation will not be adequate, and it is desirable that you select and co-register at least 20 ground control points. Some of the complications and considerations in taking this approach are detailed in the following papers:
    - Mather, P.M. (1995) Map-image registration using least-squares polynomials. *International Journal of Geographical Information Systems*, 9(5), 543-545.
    - Morad, M., Chalmers, A.I. and O'Regan, P. R. (1996) The role of root-mean-square error in geo-transformation of image in GIS. *International Journal of Geographical Information Systems*, 10(3), 347-353.
    - Unwin, D.J. & P.M.Mather (1998) Selecting and Using Ground Control Points in Image Rectification and Registration, *Geographical Systems*, 5(3), 239-260.
5. This graphical exercise is intended to help fix ideas about co-ordinate transformation in integrating digitizer data into a GIS. First study the outline of affine transformation in O'Sullivan and Unwin (2010, pages 322-324). To do the exercise, you will need some lined graph paper, some tracing paper (ideally transparent lined graph paper), a pencil and, perhaps, a calculator or spreadsheet.
    a) Create a grid on your graph paper with X and Y axes each going from 0 to 100.
    b) On this grid mark eight randomly located points and read off their (x, y) co-ordinates.
    c) Use the tracing paper to prepare an identical set of axes, but do not mark any points on it.

d) Place the transparent grid on your original one with its origin exactly on the origin of the original one and rotate it by a small known angle (say 15^{o}).
e) Next shift (translate) the origin by a known small amount, and then mark on the transparent paper the positions of your eight points.
f) Read off the co-ordinates of the eight points in this new system.

If mathematically inclined, compute and use the affine transformation matrix for this operation.

If you have the time and resources, then there is no better way to introduce the trials and tribulations of data collection than by doing it yourself in the field. This, and the next two activities, outline possible field projects appropriate to a GIS class. They can be adapted to suit almost any disciplinary background and educational level.

6. Random points survey of attribute data. Divide the class into small teams and ask each to generate a small number of randomly selected points in a small area of country they can visit in a day's field work. It helps, and introduces the notorious traveling salesman problem, if you ask them to decide on a minimum distance route that links the points. Students then visit the points, recording a series of attributes that characterize the environment in that locality. Suitable attributes might include soil type, slope and aspect, and land use. Back at base merge all these data and use them to develop, for example, estimates of the actual land use and soil type proportions, the overall frequency distributions of slope and aspect. This can be extended to use suitable statistical analyses to examine interactions between these variables such as the impact of slope on land use. Taking it further, it is very useful to have available a GIS for the same area with an 'official' version of these same attributes taken from standard map and imagery sources. There is a useful resource data base of similar projects at the GEES Resource Database: http://www.gees.ac.uk/db/

7. Surveying by leveling to establish the z co-ordinate. Although use of a simple device such as an Abney or 'Quickset' level is a long way from current professional survey practice, asking students to undertake a simple leveling transect along a clearly defined feature (such as a beach) is a good way to introduce key ideas in GIS data collection, especially measurement error and the implicit discretization introduced by selection of the stations used.
8. The third exercise is equally far from current survey practice, but has similar educational objectives. In small student teams, ask them to use simple plane table survey with basic table and tripod, alidade and ranging rods to map a small feature, such as the boundary

of a field or short reach of a river. For maximum effect, students should have to handle the errors in their survey. A useful extension is to use basic and/or differential GPS to capture co-ordinates of the same points used in the survey. Suitable basic equipment may be hard to come by in these modern times but is manufactured for educational use. You can, of course, do exactly the same exercise using any more modern surveying equipment that is appropriate and available.

9. Organize a debate around the motion that 'This house believes that geographic data should be freely available to all the citizens of country x'. See Gold *et al.* (1992), Chapter 5 at www2.glos.ac.uk/gdn/gold/ for the usual organizational details.
10. Another suitably contentious topic for debate is that of the need for mandated standards in geographic data. A suitable motion is that 'This house believes that standards in GI data are a bad idea'. A very good place to find the case for standards is at the UK Association for Geographic Information (AGI) website: www.agi.org.uk. As detailed in the text, page 214, similar lobby organization materials will be found at: www.opengeospatial.org. As an example of the problems that adherence to a standard can generate, an instructive case is the British Standard BS 7666 on street addresses. There are many others.

FURTHER READINGS

Hohl P. (ed) 1997 *GIS Data Conversion: Strategies, Techniques and Management*. Santa Fe, NM: OnWord Press.

Jones C. 1997 *Geographic Information Systems and Computer Cartography*. Reading, MA: Addison-Wesley Longman.

Lillesand T.M., Kiefer R.W. and Chipman R.W. 2003 *Remote Sensing and Image Interpretation* (5th edn). Hoboken, NJ: Wiley.

Paine D.P. and Kiser J.D. 2003 *Aerial Photography and Image Interpretation* (2nd edn). Hoboken, NJ: Wiley.

Walford N. 2002 *Geographical Data: Characteristics and Sources*. Hoboken, NJ: Wiley.

Jensen, J.R. 2007. Remote Sensing of the Environment: An Earth Resource Perspective (2nd ed.). Upper Saddle River, NJ: Prentice Hall.

RELATED READINGS

Longley P.A., Goodchild M.F., Maguire D.J. and Rhind D.W. (eds) 2005 *Geographical Information Systems: Principles, Techniques, Management and Applications* (abridged edition). Hoboken, NJ: Wiley.

31. Encoding and validating data from maps and images, I Dowman
32. Digital remotely-sensed data and their characteristics, M Barnsley
33. Using GPS for GIS data capture, A Lange and C Gilbert

Maguire D.J., Goodchild M.F. and Rhind D.W. (eds) 1991 *Geographical Information Systems: Principles and applications*. Harlow, UK: Longman (text available online at www.wiley.co.uk/gis/volumes.html).
17. GIS data capture hardware and software, M J Jackson and P A Woodsford, pp. 239-49
34. Spatial data exchange and standardization, S C Guptill, pp. 515-30

ONLINE RESOURCES

NCGIA Core Curriculum in GIScience, 2000 (www.ncgia.ucsb.edu/giscc)
- 2.9.2. Natural Resources Data (090), Peter Schut
- 2.9.2.1. Soil Data for GIS (091), Peter Schut

NCGIA Core Curriculum in GIS, 1990 (www.ncgia.ucsb.edu/pubs/core.html)
- 7. Data input
- 8. Socio-economic data
- 9. Environmental data
- 66. Database creation