5. Preparing data for GIS use

G.J. Meaden (FAO consultant, Canterbury, United Kingdom), J. Jenness (FAO consultant, Flagstaff, Arizona, United States of America), and S.Walker (NOAA Coastal Services Center, Charleston, United States of America)

Once data for GIS use have been collected via any of the methods described in Chapter 3, the data will need to undergo a number of preparatory stages before they can be effectively used. Exactly what needs doing depends upon the format in which data have been collected. Some data may be in a usable form as digital data that only needs delivering via the Internet or CD–ROM; some data may be stored on a data collection device; and other data may be in hard copy form as paper maps, tables or lists. Some GIS workers contend that getting data suitably edited and formatted may be the most important task of any GIS project. The qualitative assembling and pre-processing of data is the primary theme of this chapter, though the chapter also describes how data are best modelled and structured for GIS use, data storage, database functioning and metadata. Having data stored in the correct format for any GIS software is fundamental to all operations. Thus, there will be strict rules on matters such as how a georeference should be stated, the range of values that are permissible, the number of decimal places to be recorded and the coding that can be recognized. Some GIS handle several data formats and there are now applications that switch data between formats.

The primary preparatory data need is that of getting data into the system in a digital format. The first method for doing this, via the use of a scanner and digitizer, has been described in Chapter 2. Here, it is simply necessary to note that good data presentation from the scanning process is best achieved through the use of high-quality maps and through careful geo-registering of maps, plus accurate recording of associated information on colour, attribute, line thickness, etc. Successful digitizing, either via a table or via the "on-screen" method, is only achieved with careful attention to detail and accuracy in the line following process, i.e. to avoid duplicated lines or lines that do not properly complete polygons. The results from the digitizing process should be stored in individual thematic files, such as for coastlines, railways, rivers and streams and bathymetry. As well as scanning and digitizing, hard copy tabular data can be input manually through the creation of files (typically in database or spreadsheet formats) using computer keyboard entry. To be of use in any GIS, manually entered data must have suitable georeferences or coded references to place names, zip codes, land use, etc. The final method of data entry is through the direct transfer of digital data via CD–ROMs, DVDs, memory sticks, data loggers or the Internet. Because much of these data will not have been collected with any specific GIS project in mind, care must be taken with respect to their date and resolution and to the classification categories

or data formats used, among other things. There might also be a problem with respect to the size of data files, especially if using sonar or remote sensing data, so adequate data storage capacity must be available.

Once the required digital data have been assembled, it will be essential to validate the data for correctness and for being up to date. Any mistakes or changes that are necessary can then be made through various editing processes after which the data can be suitably organized and stored. Maintaining up-to-date edited data can be a costly and time-consuming task, this being a function of numerous sources of error such as inexact digitizing, unsuitable formatting, incorrect data recording and uncertainties in data classification. Most GIS have their own editing function tools. There are two main types of editing:

- **Graphical editing.** This refers to making corrections to any points, lines or polygons (graphical entities) on a digital map. The need for graphical editing mainly arises from incorrect digitizing, and typical errors are described in the technical paper. Most GIS software contains tools that help identify errors and then edit any digitizing errors.
- **Non-graphical editing.** Here, the concern is with correcting the attribute data that corresponds to mapped features, e.g. place names or feature types. Sources of errors mainly include incorrect feature classifications or poor keyboard entry. Non-graphical errors are more difficult to detect and it may be impossible to check every feature. However, Heywood, Cornelius and Carver (2006) show that data can often be verified by seeking out impossible or extreme values, by looking for inconsistencies, and by using scattergrams or trend surface analyses to identify data anomalies.

Editing will be a continual GIS prerequisite because new data sets are always being added and existing data sets will undergo many revisions. The need for data validation and editing will be especially important to fisheries GIS work because the nature of many spatial distributions in aquatic environments is both temporary and volatile.

For maximum efficiency, all data collected must be appropriately stored, organized, managed and shared, and this is accomplished via the use of files, databases and database management systems (DBMS). Files are the basic form of data storage used in any computing environment. A file is usually a single collection of records about a certain theme, and it may cover a certain time period and a specific location or area. Files themselves may take several forms, such as text, tables, imagery or shapefiles, and the data might have been entered onto a spreadsheet or into a database or stored within a file format that is specific to a particular GIS. Within any file, all data must be consistently stored using the same format throughout. A collection of files is usually referred to as a database, and this may cover a wider set of themes or areas than individual files. For use in a GIS, it is essential that all the data (held in files within the database) can be linked to some form of georeference, i.e. either geo-coordinates or names given to particular areas. This latter georeference is called a "unique identifier". A single database may be very large and may cover numerous related themes, plus all the attributes

related to each theme, e.g. perhaps a lake fisheries database would have files on physical characteristics of the lake, species in the lake, management policies for the lake, land use around the lake edge, etc., with each file storing information on the different attributes pertaining to each theme. Databases are important to GIS software because they provide structure to the stored data allowing the data to be manipulated in a multitude of ways. Within a GIS or IT department, a series of databases may be held on one computer, although larger departments may have specialist server computers that supply data to all users having access rights.

In order to attain wide functionality, collections of databases within a workplace are managed by the DBMS. In a sense, it is the DBMS that lies at the core of GIS capability. DBMS software provide the ability to load, access, modify and maintain data, plus a range of security, back-up and administrative tasks; most GIS software now contain their own internal DBMS. But unlike most DBMS, GIS must be capable of handling spatial demands of the data, allowing the GIS to be manipulated, to be queried or searched, and generally to maximize the efficiency in which the data can be utilized for carrying out the whole range of GIS tasks. Although the DBMS can be structured in a variety of ways, the so-called relational DBMS model has become the favoured model for GIS use. The technical paper provides details on the working requisites of this DBMS model and provides a worked example of how files can be joined and manipulated using relational DBMS techniques.

For users of large quantities of data, it becomes increasingly important to keep a record about all of their data sets. This record is known as the metadata, a collection of which may form a metadatabase. Metadata include a range of essential facts about the data sets, including information on data quality, data ownership, themes or variables included, coordinate systems used, data exchange formats, codings used for attributes, and organizations producing the data. Having information on these factors ensures that data can be used with confidence, and it will provide clues as to when data may need updating. Many database managers will not add data sets to their databases if the sets do not come with their own metadata.

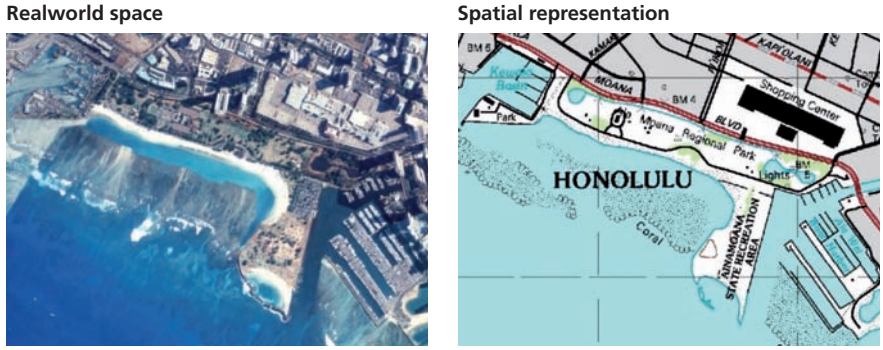
With data now being stored and ready for use, attention turns to how best to show the real world in a mapping format, i.e. it is important that all features of the real world that need to be mapped for any given GIS project be displayed in a meaningful way. To achieve this, it will be necessary to identify classes of features, e.g. buildings, roads, land uses and sediment types, and for each class and subcategory it will be necessary to make a decision on how features might best be displayed in a mapped form. This cartographic classification will require much generalization, that is, the simplification of the real world by omitting unimportant features, by simplifying real-world shapes, and by the features classification process itself (see Chapter 7). The aim of this real-world modelling process is to achieve a set of symbols that best and most accurately create a visually pleasing, intuitive or understandable mapped representation for the area of interest, plus the means of portraying the non-graphical entities on the maps (place names or

features types). Figure 5 illustrates how real-world features (as detected from a remotely sensed image) might be represented as mapped data. It will be clear that the way in which features are displayed on a map will be scale dependent, i.e. at a small-scale (large area) towns might simply be shown as a dot, whereas at a large scale (small area) a town will consist of roads, buildings, green spaces, etc. A close inspection of the graphical symbols that depict features in any mapped area show that they can all be classed as being either points, lines or polygons, and, through the use of some kind of georeferencing, these are convenient formats in which information about map features can be stored in a database, such that any feature can then be digitally mapped. GIS software has the capacity to use the graphical data to produce maps that are either structured in the vector data format or in the raster data format.

The vector data format relies on having the exact geo-coordinates for every point, line or polygon needing to be mapped. To map any point, single x and y coordinates are necessary so that the GIS can accurately position the point in relation to a mapped project area. Straight lines can be drawn by knowing the start and end coordinates for the line, after which the GIS simply joins these two points. Curved lines are either drawn by knowing the coordinates of numerous coordinates around the curve and then joining them up, or they can be drawn by using an algorithm embedded in the software that relies on the use of selected coordinates around a curve to plot the likely position of the curve. Polygons are just lines that start and end at the same geo-coordinate, i.e. to completely enclose a mapped area. The technical paper describes in more detail various specialized vector file formats. The vector format is particularly suited to drawing accurate linear versions of maps and it also allows the GIS to perform specific calculations efficiently based upon linear lengths or areas.

The raster data format relies on the mapped area being comprised of equal size cells (or pixels). These cells are usually square shaped, but theoretically they could be rectangles, equilateral triangles or hexagons. Individual cells in a raster format may be identified by sequential alpha and/or numeric figures for the

FIGURE 5
Real world spatial features represented as geospatial (mapped) data



Source: Feinholz and Walker (2007).

columns (x axis) and for the rows (y axis). Raster mapping information is built up in layers, each representing a different theme such as land use, sediment type and water temperature. Each raster cell is allocated a single code (or value) based on the predominant feature or class occupying that cell. The coding given to each cell might be in terms of a numerical value, a weighting, a coding allocated to a colour, a reflectance value, etc. The dimension of cells is important to working in a raster format because larger cells occupy larger portions of the real-world surface, and they therefore produce a cruder or more generalized map; conversely, smaller cell sizes produce more accurate maps, but they require more data storage space. The use of very small pixels allows maps to be produced that look the same as vector drawn maps. Because of the large amount of data storage associated with the raster format, there are various methods for data compression that are described in the technical paper. In the past, it was important to select to work in either vector or raster format, but today most GIS can convert between working in either of the formats.

Two other data models are important to GIS use, and they are concerned with the structuring of data for: (i) 2.5D terrain modelling; and (ii) network models. Terrain modelling in 2.5D incorporates the x, y and z axes, but with the z axis only referring to the ground height above mean sea level (or an agreed datum line) or the water depth below a datum line. Terrain modelling itself is conventionally subclassified into one of two data formats according to whether they are using vector or raster formats, i.e. either triangulated irregular networks (TINs) that use the vector format or digital elevation models (DEMs) that use the raster format. It can be envisaged that the ability to incorporate the z axis into fisheries GIS work considerably broadens the potential for spatial analyses, i.e. because fisheries and mariculture function almost entirely in a 3D environment. The functioning of GIS-based terrain modelling is discussed in Chapter 7.

A network is an interconnected set of points, lines and polygons that represent possible routes or "pathways" from one location to another, and networks may represent natural features such as waterways, migration routes and vegetation corridors or, more frequently, human engineered structures such as roads, railways, air routes, fencing, pipelines, cables and boundaries. Given the wide array of networks, their modelling is an important function for GIS; for example, it is likely that most aquaculture facilities will need to include network analyses as a fundamental input to their location decision. Network modelling is usually performed using the vector data format. For modelling purposes, networks can be conveniently conceived in terms of links and nodes, with the former being the routeways themselves and the latter being junctions or start and end locations. Values can be assigned to links in terms of distances, travel times, fuel used, financial costs, etc., and to nodes in terms of population, number of businesses, tourist attractions, etc. The functional GIS-based analyses performed using network modelling is described in Chapter 7.