

Data Integration

Data integration is the process of combining data of different themes, content, scale or spatial extent, projections, acquisition methods, formats, schema, or even levels of uncertainty, so that they can be understood and analyzed. There is often a common display method used with integrated datasets that, although they are not fully processed, allows information to be passed between them. Integrating different types of data in a GIS often provides more information and insight than can be obtained by considering each type of data separately. It also aids in the detection and removal of redundancies or inaccuracies in the data (in both location and attribute). The layer stack concept that is so illustrative of GIS (often implying the overlay of maps) helps one to understand data integration in a vertical sense. It will often take place in a horizontal sense also, such as the matching together of adjacent map edges. Data integration is one of the main reasons why GIS software is used, and must often take place before spatial analysis can be performed on the data.

Principles and Practices of Data Integration

Data integration often starts with the compiling of various datasets from different sources and at varying scales, formats, and quality, or with acquiring the data in the field (i.e., the accurate sensing and collection of measurements from the environment) and the transformation of these measurements from raw to fully processed for GIS input and analysis. Some of these datasets are already in GIS format and have metadata (or descriptive information about the data) associated with them. Other datasets are not in

GIS format. Therefore in data integration, not only do a wide variety of data sources need to be dealt with, but a myriad of data “structures.” For example, the user may need to integrate chemical concentrations stored as either spreadsheet tables, database management system files, or text files, with satellite images, with gridded topography or bathymetry. These must all be converted to a form that a GIS will accept, and the accompanying metadata must be created where necessary.

After all datasets have been converted into a common GIS format, the next step is to load all of the data into the GIS, which will often require the aid of simple data import routines within the GIS. Once the datasets are in the GIS, they must be further manipulated so that they all register, or fit together, in space, and where possible, in time. The data integration process may need to deal with scales of information ranging from hundreds of kilometers to millimeters, and decades to milliseconds. Here the rendering of GIS files to a common spatial reference system (usually comprised of a common datum, map projection and spatial extent) is very important. Although datasets are in a common GIS format (such as a shapefile, coverage, or geodatabase), they may have been derived from different map projections, and must therefore be converted to a common map projection.

Some consider an additional step in data integration to be the presentation and analysis of the integrated data, because once the data are converted to a single reference system, different categories of data can be represented by interoperable thematic layers, and the relationships between these layers can be easily established, either by looking at them in map form (e.g., turning the layers on and off) or by performing a series of spatial overlay operations. The results of these operations may then be input into either graphic

or statistical analysis routines. These graphical, relational, and statistical associations between the integrated data elements may all be used to infer the corresponding relationships that actually exist in the natural environment,

Examples of Data Integration Projects

Excellent examples of data integration projects are too numerous to mention here, but two offer good illustrations. The primary goals of the U.S. Geological Survey's National Map are to provide a consistent framework for the geography of the entire U.S., as well as to allow public access to high quality geospatial data and information from multiple sources (data partners). It seeks to integrate the foundation and framework layers of orthographic imagery, land terrain/elevation, boundaries, transportation, hydrology, land cover, and geographic place names, with a myriad of specialized data sources from their partners at all levels of government, various Indian tribes, academic institutions, and non-governmental organizations. This is indeed a huge data integration challenge.

A more general example is the data integration challenge provided by research in the field of oceanography where the integration of multidisciplinary data gathered from many different kinds of instruments is of great importance. Here marine geologists, chemists, biologists and physicists must often work together in order to understand the "bigger picture" of natural ocean process such as seafloor spreading, ocean-atmosphere interactions, tracking and modeling of El Niño, mapping of global weather patterns, or the determination of various biophysical properties of the oceans (i.e., temperature, chlorophyll pigments, suspended sediment, and salinity). In oceanography, the cost of acquiring the data alone (e.g., an oceanographic research vessel usually costs over

\$25,000 a day to operate) justifies the development of dedicated systems for the integration of these data.

The introduction of a wide range of sophisticated vehicles and instruments for surveying the ocean has necessitated the development of reliable data integration procedures for the various data streams. For example, bathymetric data from a swath mapping system located *underneath* a ship may need to be georeferenced to underwater video images or sidescan sonar data collected from a vehicle towed *behind* the ship and several meters *above* the ocean floor, to sample sites, observations, temperature measurements, etc. collected from a submersible or remotely-operated vehicle (ROV) launched *away from* the ship and operating directly *on* the ocean floor, to earthquake data obtained from an ocean bottom seismometer anchored on the sea floor. The integration of remotely sensed images from space, with *in situ* data (i.e., point, line, are polygonal data gathered “on site,” at sea) is also an important consideration. The data produced by all of these different sensors will invariably have different dimensionalities, resolutions and accuracies.

As transmission rates of up to several gigabytes per day at sea become more and more commonplace, the ability to assess in real-time ocean floor data collected at these different scales, in varying formats, and in relation to data from other disciplines has become crucial. Here GIS is of critical importance as it fulfills not only the requirement of rapid and efficient data integration but also of combining or overlaying data of the same dimensionality to facilitate scientific interpretation of the data. This also serves as an efficient means of assessing the quality of data produced by one instrument as compared to another. As a result, applications of GIS for ocean mapping and a wide

range of environmental fields have progressed from mere collection and display of data to the development of new analytical methods and concepts complex simulation, modeling.

Related Issues

The growth in information technology has led to an explosion in the amount of information that is available to researchers in many fields. This is particularly the case in the marine environment where a state of the art "visual presence" (through real time video or 35 mm photography) may result in the acquisition of data that quickly overcomes the speed at which the data can be interpreted. The paradox is that as the amount of potentially useful and important data grows, it becomes increasingly difficult to know what data exist, where the data are located (particularly when navigating at sea with no "landmarks"), and how the data can or should be accessed. In striving to manage this ever-increasing amount of data and to facilitate their effective and efficient use, metadata becomes an urgent issue in effective data integration.

Geographic information scientists have addressed many research topics related to data integration, such as the management of very large spatial databases and associated uncertainty and error propagation, the designation of "core" or "framework" datasets which form the base datasets in any integrated collection, and the development of standards for spatial data and metadata. Metadata should be created in compliance with a standard such as that created by the Federal Geographic Data Committee (FGDC) or the International Standards Organization (ISO). Protocols and maintenance procedures for data contributed to archives, clearinghouses, or other distribution points should also be

documented, as well as policies and procedures for future data acquisitions. Geographic information scientists also devise appropriate data quality criteria such as the development of relative measures of quality based on positional differences in datasets.

Also related to data integration is the concept of data lineage, or the history of how the spatial data were derived and manipulated. The U.S. National Committee for Digital Cartographic Data Standards has defined lineage as information describing source materials and the transformations used to derive final digital cartographic data files. A report of lineage is therefore intended to serve as a communication mechanism between the data producer and the user, a kind of "truth in labeling" statement regarding the process leading up to the present state and quality of GIS-derived products. For example, if a metadata record includes ancillary information such as sensor calibration, data quality assessment, processing algorithm used, etc., the lineage includes the time stamp for each of these, and information on the manipulations performed on the dataset since it was initially created. Because oceanographic and other kinds of environmental data often come from a variety of sensors, differing in resolution and covering different geographical areas, lineage documentation is especially important for assessing data quality, data history, and error propagation.

The fact that datasets have been routinely collected at different times is a further consideration. The most recent dataset is usually assumed to be the most "correct," provided that no special error conditions are known to have affected the sensor. In practice, small variations in time within or between datasets gathered at sea are often ignored to simplify the analyses and modeling. A prime consideration as researchers, managers, organizations, or individuals seek to integrate and maintain data, will be to

always provide information on the source of data input to the GIS, as well as database and cartographic transformations performed on the data within the GIS, and on resulting input/output relationships between source-, derived-, and product-GIS data layers.

Dawn J. Wright

See also Data Conversion; Edge Matching; Error Propagation; Framework Data; Metadata; Projections; Spatial Analysis; Spatial Data Infrastructure

Further Readings

Bonham-Carter, G.F. (1991). Integration of geoscientific data using GIS. In D.J. Maguire, M. F. Goodchild, & D. W. Rhind (Eds.) *Geographical Information Systems: Principles and Applications* (pp. 171–184). New York: John Wiley and Sons.

Jensen, J., Saalfeld, A., Broome, F., Cowen, D., Price, K., Ramsey, D., Lapine, L., & Userly, E. L. (2004). Spatial data acquisition and integration. In R. E. McMaster and E. L. Userly (Eds.), *A Research Agenda for Geographic Information Science* (pp. 17–60). Boca Raton, FL: Taylor & Francis-CRC Press.

Flowerdew, R. (1991). Spatial data integration. In D.J. Maguire, M. F. Goodchild, & D. W. Rhind (Eds.) *Geographical Information Systems: Principles and Applications* (pp. 375–387). New York: John Wiley and Sons. [Also Online]. Retrieved August 16, 2006 from <http://www.wiley.com/legacy/wileychi/gis/volumes.html>

Wilkinson, G.G. (1996). A review of current issues in the integration of GIS and remote sensing data, *International Journal Geographical Information Science*, 10, 85–101.