4.3

GEOSPATIAL DATA IN EMERGENCIES

MICHAEL F. GOODCHILD

IN RECENT YEARS, roughly dating from the popularizing of the Internet beginning in 1993, there has been rapid and massive growth in the use of electronic networks for sharing geospatial data. Today, it is possible to find a vast resource of geospatial data, along with such derivative products as maps, all distributed over the tens of millions of servers connected to the Internet, and accessed using simple and widely available tools. Many governments have developed digital clearinghouses and warehouses of geospatial data as part of efforts to sponsor and build spatial data infrastructures (see, for example, the National Geospatial Data Clearinghouse of the Federal Geographic Data Committee, and its proposed Geospatial Data One-Stop, http://www.fgdc.gov). This process is evident at many scales, from cities and counties to states, nations, and the globe (Masser 1998, National Research Council 1993, Rhind 1997). The totality of geospatial data resources available through some servers exceeds one terabyte (TB), and one might guess that the total global geospatial data resource available in digital form through the network is now of the order of one petabyte (10^{15} bytes).

DATA ACCESS AND VALUE

Such data resources are of enormous potential benefit in applications relating to terrorism. Many unexpected events, such as the bombing of the Murrah Building in downtown Oklahoma City and the destruction of the World Trade Center, are associated with precisely defined locations on the Earth's surface. In the immediate aftermath of a disaster event, it is necessary for those responsible for recovery to assemble and provide rapid access to information on building plans, local streets, facilities such as hospitals, the local distribution of daytime

and residential population, utility corridors, and many other types of information, many of which are geospatial. The key to all of this information is the location of the specific event or impact area. Since such events are virtually impossible to anticipate and can occur in an infinite number of locations, it is essential that the means exist to search, retrieve, and assemble geospatial data based on the geographic locale (or key) as rapidly as possible, within seconds or at worst minutes. The ability to search based on geographic location defines a special type of library, or *geolibrary* (National Research Council 1999).

Unfortunately, and as a result of a number of early and fundamental design decisions in the development of geographic information systems (GIS), the Internet, and the World Wide Web (WWW), would-be users of the distributed resource of geospatial data are faced with daunting problems. There are few effective catalogs of the resources available for browsing and supporting searches for specific data. Once found, data may be very difficult to integrate because of incompatible formats and inaccuracies. Easing these problems is important for all users, but perhaps most important in an emergency, when time is of the essence. This paper explores options for improving access to geospatial data in emergency situations. The following sections expand on each of the major issues—support for search, interoperability of formats, and limitations on accuracy—and assesses the prospects for progress. The final section discusses some remaining issues and prospects for the future.

FINDING DATA

The Internet exhibits what one might term *functional* organization, rather than *spatial* organization, in the sense that its design attempts to ignore distance, and to provide equal access to information independent of location. The IP (Internet protocol) addresses used by the network resolve geographic location only coarsely, if at all, and the time required to access and retrieve information is virtually constant however great the distance between the data host and the user. The system provides no means of identifying servers that are geographically close, or of searching only over servers in a given geographic area. Instead, search capabilities are provided by a number of search engines such as Altavista or Google, using catalogs of Internet-based information resources built largely automatically through the use of *crawlers*, programs that traverse the Internet's servers following the hyperlinks of the WWW, finding and then extracting significant keywords. These keywords are then sorted into an index or catalog, which can be accessed by a user interested in a particular topic.

Unfortunately keywords are not a good way of identifying the presence of geospatial data, or of detecting those specific properties that would be of interest to a user. Although search engines provide an excellent mechanism for finding information expressed in the form of text, they are not an effective basis for searching or browsing the distributed resource of geospatial data. Other tools are more successful, but over more limited domains. For example, the MapFusion software developed by Global Geomatics Inc. (http://www.globalgeo.com) is able to scan any file system, either local to the user's system or distributed over a defined Internet domain, and detect files that use any of a wide range of standard GIS formats (Goodchild 2002). The necessary characteristics, such as scale, authorship, and date, are extracted from the file's own header information provided the GIS format contains such metadata (data about data). But, such systems are not likely to scale to the magnitude of the Internet, and would have no way of accessing or scanning most file systems (unlike WWW pages, which are more easily accessed from remote systems).

While some progress has been made in recent years in building WWW crawlers specifically designed to detect and catalog geospatial data, the problem remains a serious impediment to the successful use of distributed data resources. Instead, many agencies and companies have developed other approaches, the most common of which is the clearinghouse or portal. A clearinghouse is a WWW site containing geospatial data, together with a catalog that allows a user, having reached the site, to search for data meeting specific requirements. In some cases the clearinghouse is a warehouse, maintaining all data locally; in other cases it is a portal to a distributed resource, and several methods exist for registering each of the distributed resources with the portal (such as the ESRI Geography Network, http://www.geography network.com). None are fully automated, however, for the reasons discussed above.

Two problems limit the success of this approach. First, no single site can possibly succeed in establishing a monopoly on access to geospatial data. Different levels of government with overlapping jurisdictions frequently vie for this role, and compete with the private sector, and with individuals. Second, since it follows from the previous point that there will always be more than one clearinghouse, it is necessary for the user to possess some method of knowing where to look for a given set of data; in other words, to possess collection-level metadata that describe the general characteristics of the contents of any clearinghouse. At this point no such mechanisms exist, so users are forced to rely on personal knowledge, interpersonal networks, and guesswork based on simple heuristics (for example, that a data set is most likely

to be found on a server maintained by a governmental agency whose jurisdiction most closely matches the geographic coverage of the data set) (Goodchild 1997).

INTEROPERABILITY

The technology of GIS has developed over the past four decades largely in the absence of strong overarching theory, and as a result each vendor of GIS software has tended to adopt a distinct terminology, and distinct data formats. While terms such as *topology*, *layer*, and *coverage* are widely used in the GIS industry, their meanings are constrained only by the comparatively vague limits of intuition, rather than by formal theory (topology, for example, has a formal mathematical meaning that has drifted substantially in GIS usage). Many standards have emerged, but usually in narrowly defined communities such as defense, or civilian government agencies, and each new effort to standardize seems merely to add to the already long list of formats. *Interoperability* is defined as the ability of systems to exchange information, based on shared understanding of meaning (*semantic* interoperability) and mutually agreed formats (*syntactic* interoperability) (Goodchild et al. 1999).

Lack of interoperability has created a very significant paradox for GIS. On the one hand GIS is a technology able to analyze vast amounts of information at close to the speed of light. On the other hand, when data are assembled from distributed sources it is almost always necessary to spend a large amount of time reformatting, interpreting different meanings, reconciling differences in classification schemes, and forcing consistency in map projections and coordinate systems. It often can take months to prepare data for just a few seconds of analysis.

In recent years the Open GIS Consortium (OGC) (http://www.opengis.org) has emerged as a powerful force in the achievement of greater interoperability. OGC's approach is not to force standardization, which itself requires enormous investment, but to use information technology and general specifications to overcome differences. For example, servers operating in compliance with OGC's WWW mapping specifications are able to supply users with data using a general format (XML or Extensible Markup Language) that is readily compatible not only with OGC's own internal formats, but with the user's client system. These new approaches have been demonstrated very successfully, in applications ranging from environmental monitoring to emergency response (for examples see the OGC Website).

INACCURACY

Perhaps the most problematic issue facing users of distributed geospatial data resources is data inaccuracy. In its most obvious form, data inaccuracies manifest themselves when an user overlays two data sets, from different servers, with different origins, and finds that they fail to fit perfectly. Since it is impossible to measure location on the Earth's surface perfectly, any two data sets will always fail to fit at some scale, but unfortunately the scale at which the problem becomes evident, and impacts applications, is often surprisingly coarse. For example, the widely used data sets of street centerlines, many of them derived from the Bureau of the Census TIGER files, have positional accuracies ranging from 10 meters to 50 meters, sufficient to confuse two ramps at a complex freeway junction, or to mix up two parcels in an urban area. The practical consequences become inescapable when such data sets are used to dispatch emergency vehicles in response to calls based on GPS locations, which themselves may be inaccurate by as much as 10 meters or more.

This problem is not likely to be resolved easily, and is only part of a much larger issue related to the certification of the accuracy of geospatial data. At best, the accuracy of geospatial data sets is described by the creators of the data sets as an important part of metadata, and distributed to users automatically. At worst, however, there is no obvious source of knowledge on data quality for many of the geospatial data sets currently obtainable via the WWW. Positional accuracy is traditionally related to map scale, or the representative fraction that proportions distance on the map to distance on the ground. On maps at 1:250,000, positional errors of as much as 100 meters are acceptable according to national map accuracy standards. But GIS and the WWW make it comparatively easy to integrate data derived from such relatively coarse and inaccurate maps with data from much more accurate sources, such as records of property ownership or engineering-grade surveys of infrastructure.

CONCLUDING COMMENTS

Geographic information technologies such as GIS, combined with the power of the Internet for rapid sharing of information, create an exciting range of possibilities for those charged with anticipating and responding to terrorist acts. Geographic location is the key attribute used to define a search for relevant information, and geospatial data clearly are extremely valuable for a host of applications related to disasters, hazard vulnerability and response.

The vast resources of geospatial data available through the Internet and WWW coupled with expertise in the use of GIS, provide a powerful basis for addressing terrorism. But this optimism must be tempered by knowledge of several critical issues related to data access and use. This paper has identified three: a lack of efficient mechanisms for searching over a distributed data resource for appropriate geospatial data; a lack of interoperability between different data sets; and inaccuracy in data. Another key issue related to use of geospatial data in emergencies, the development of common data models and procedures, was addressed elsewhere in the first essay of this chapter. The research community is actively pursuing all of these, and organizations such as the Federal Geographic Data Committee, the Open GIS Consortium, and the International Standards Organization are making substantial strides in improving interoperability through common specifications and standards. While there is progress, much more needs to be done if the current barriers to effective geospatial data access and sharing are to be overcome.