4

Different Data Sources and Diverse Data Structures: Metadata and Other Solutions

Michael F Goodchild

Summary

Geocomputation almost by definition requires access to large quantities of geographical data, and it is increasingly common for such data to be supplied using technologies that support search and retrieval over distributed archives, such as the World Wide Web. It is essential therefore that it be possible to define the characteristics of needed data; to search for suitable sources among archives scattered over a potentially vast distributed network; to evaluate the fitness of a given dataset for use; and to retrieve it successfully. These stages require the development of an array of tools, and associated standards and protocols. The term 'metadata' is commonly used to refer to languages designed for the description of the contents of a dataset, to facilitate its discovery and evaluation by a search engine, as well as its successful transmission and opening by the user's application. In the area of geographical data, the most widely known metadata standards are the Content Standards for Digital Geospatial Metadata, developed and implemented by the US Federal Geographic Data Committee.

4.1 Introduction

Geocomputation has a large appetite for data. While the literature on cellular automata and artificial life shows that it is possible to build interesting simulations within an undifferentiated spatial frame with virtually no input except for model

Geocomputation: A Primer. Edited by Paul A Longley, Sue M Brooks, Rachael McDonnell and Bill Macmillan.
© 1998 John Wiley & Sons Ltd.

parameters, geocomputation focuses on modelling processes on geographical landscapes that can be sharply differentiated. Its processes respond to assorted boundary and initial conditions, and these must be represented therefore by input of appropriate geographical data. The parameters of geocomputational models may also be spatially variable, and must be represented with potentially extensive input data. In both of these cases the data serve to differentiate the geographical landscape, and are therefore geographical in the traditional sense, representing the variation of conditions or attributes over geographical space: in general, $f(x,y)$, where x and y are positional variables and f is an attribute. Geocomputation is similarly a heavy producer of data, and requires tools for the analysis and display of voluminous simulation results.

The geographical landscape is inherently complex, since x and y define a continuous frame, and its representation can in principle require an infinite amount of information. But the processes being modelled in geocomputation are likely to have inherent *scale*, meaning that there exists some linear measure P that is a property of the process, such that variation over distances less than P has effectively no impact on the outcome of the process, and therefore need not be input (or *resolved*, to use a term common in the modelling community). The literature on physical environmental processes contains many attempts to define P both approximately and precisely for specific processes (see, for example, Delcourt et al 1983; Rosswall et al 1988; Ehleringer and Field 1993).

Although the literature of geographical information systems (GIS) has recently begun to include discussion of many types of data, including multimedia (see, for example, Batty et al, this volume; Clarke, this volume; Raper 1997), it seems reasonable to assume that in almost all cases the input to geocomputational models will be strictly geographical, and therefore will follow one of three classes of conceptual data models: discrete objects, fields, or their equivalents on 1-dimensional networks embedded in two or three spatial dimensions (Goodchild 1992). This raises one specific issue, however: the set of standard GIS data models does not include all of those data models commonly used in numerical modelling in space. Specifically, while finite-difference modelling (see Burrough, this volume) uses discrete spatial elements that are recognisable as raster data models in GIS, finite-element modelling (FEM) makes use of grids or *meshes* that do not have recognisable equivalents in mainstream GIS software. GIS has yet to recognise the importance of representing a field using quadratic functions over triangular elements (GIS TIN implementations use only linear functions), or of polynomial functions over quadrilateral elements (for a review of FEM mesh techniques see, for example, George 1991; Knupp and Steinberg 1993).

Implicit in much discussion of geographical data is the notion of *sharing* (Onsrud and Rushton 1995). Many types of geographical data are collected for very broadly defined purposes, and are widely disseminated and used. A distinction is often made between *framework* data and other types (NRC 1995): framework geographical data are defined as having general use for the purposes of positioning, and for construction of other, more specialised data that can be referenced to them. Many other types of geographical data are also collected for diverse uses, and the processes of collection and use of data can be widely separated geographically, in

time, and by discipline – for example, the information contained in a soil database may be collected by a soil scientist, but used by a meteorologist in a model of atmosphere–soil moisture transfer. Remote sensing is a major source of geographical data (Curran et al, this volume), and here also the funding, design and construction of the sensor may have little direct connection with the data's eventual use.

A complex set of arrangements has evolved for production and dissemination of geographical data, and forms the data supply context for much of geocomputation. Recently this system has been revolutionised by the arrival of the Internet and the World Wide Web (WWW), which have removed almost entirely the costs and delays associated with traditional dissemination methods. In this new world data are to be found in widely distributed archives, ranging in size from personal servers built by individuals to make small datasets available to colleagues, to the massive servers maintained by the US Geological Survey's Eros Data Center, or the US National Aeronautics and Space Administration's EOS-DIS (Earth Observing System Data and Information System) for dissemination of vast amounts of Earth imagery and other data. To find data in this loosely coordinated and vastly complex environment the user needing data for a specific purpose must somehow:

1. *Specify* that need in terms whose meaning is widely understood.
2. Initiate a systematic process of *search*.
3. *Assess* the suitability for use of any item identified as potentially useful by the search process.
4. *Retrieve* the data using available communication channels.
5. *Open* the data for use by a local application.

This new framework differs markedly from its traditional precursor, which relied extensively on individual expertise and assistance. In most cases the potential user of data would have been a *spatially aware professional* (SAP) with knowledge of a specialised vocabulary shared with other SAPs. He or she would have interacted with a custodian of the data, perhaps at a map library or in a government office, and the telephone number of the custodian or a previous user may have been entirely sufficient to provide the necessary information about the data in question. Data would have been supplied on tape, perhaps by mailing, or in hard copy form to be digitised by the user. Much time would have been spent making the data compatible with the local application, perhaps by reformatting. Much of the available data would have been produced centrally, by a government department funded at public expense, whereas today data are increasingly available from individuals, or local agencies. With central domination of production, it was possible for uniform standards to be imposed; today, a plethora of standards have emerged as a result of marketplace competition and local autonomy.

In short, geocomputation, with its extensive data demands, is arriving as a novel paradigm at a time when many traditional arrangements for production and dissemination of geographical data are breaking down, and are being replaced by a much more flexible, localised, autonomous, and chaotic system that is at the same time much richer, with far more to offer. While new technology has made far more

data available, it has also created massive problems in making effective use of its potential. Paradoxically, only the technology itself can provide the basis of solutions. The purpose of this chapter is to examine efforts to deal with these issues, and specifically to provide tools for tackling the five stages identified above.

The next section of the chapter reviews the traditional approach to these issues, using as a framework the services provided by the research library. This is followed by five sections, one on each of the five stages of data acquisition. The chapter ends with concluding comments. Much of the discussion is based on the author's experience with the Alexandria Digital Library (Smith et al 1996), a project to make a large, distributed resource of geographical information accessible via the Internet.

4.1.1 The Library Service Model

Libraries have existed for centuries, and one of their purposes has been to satisfy the types of needs identified above, by adopting a very general approach to information retrieval. Libraries help users specify needs by providing structured tools; a thesaurus, for example, allows a user to translate terms into those accepted by the library as the basis of its own information abstraction and cataloguing. Libraries support search by abstracting information about every information object (book, journal article, or map) using standard formats. Assistance is available to the user as he or she searches for suitable objects, and assesses their fitness for use, and the user is able to browse through many information objects in searching for the best fit to a requirement because information objects are typically shelved by subject. The retrieval of information objects is made possible by assigning them unique codes. Only in the last stage, the opening of data for use by an application, is there no direct analogue among traditional library services.

While the library service model appears suited to any type of data, in practice it has not dominated the dissemination of geographical data, and alternative arrangements have emerged that are largely outside the library paradigm. Geographical data have been difficult to catalogue and abstract, cumbersome to store, and consumed by a comparatively small and specialised community. As the previous arguments and the next sections demonstrate, these assumptions are increasingly untenable, and geographical data are increasingly regarded as part of the information mainstream.

4.2 Specification

Consider the archetypical application of geocomputation. A user needs to model processes in a given geographical area, and requires data that can specify initial conditions, boundary conditions, or the variation of parameters across the area. The *footprint* of the study area is thus the most important characteristic of data, and the primary basis of search. Footprints can be defined in two ways: by specifying the bounding coordinates, or as one or more place-names. A *gazetteer* provides the ability to translate between the two options, but unfortunately few place-names have

well-defined footprints, gazetteers rarely provide more than a point reference, and only certain types of place-names appear in gazetteers. Thus bounding coordinates are clearly preferable to place-names as a rigorous basis for defining both the requirements of a project and the coverage of available information objects.

Unfortunately the traditional library has no analogue of a search that is driven by a set of bounding coordinates. Information objects in the library are classified by subject, using a discrete and finite set of topics in a controlled vocabulary. Information objects are also catalogued alphabetically by author, and alphabetically by title, but in both cases the space is one-dimensional, discrete and finite. A search based on footprints is two-dimensional, continuous and infinite, and clearly not supportable using traditional library techniques, which is one reason why map libraries have been so difficult to catalogue.

Let A denote the footprint of the project, or the specification of the geographical coverage aspect of the requirement. Let B_i denote the footprint of a geographical-information-bearing object (GIBO) i. Assume that both A and B_i are defined as rectangles aligned with latitude and longitude. Although some precision is lost, the benefits of this assumption in improved performance would seem to far outweigh the disadvantages. The goodness of fit of B_i to the specification A can be measured in various ways. A simple Boolean search might require that A be wholly contained within B_i, but this would imply that all GIBOs covering the entire surface of the Earth are perfect fits to all specifications. More useful is the measure $\|A \cap B_i\| / (\|A\| \|B_i\|)^{1/2}$, or the area of intersection divided by the square root of the product of the areas. This measure is 1 if the GIBO's footprint fits the specification perfectly, and decreases if either the GIBO only covers part of the specification footprint, or the specification covers only part of the GIBO footprint. Goodchild et al (1998a) have generalised this to the case of fuzzy footprints, where the footprint of either the GIBO or the specification is uncertain or poorly defined.

After location, the specification's next most important components are likely to be theme, date and level of detail. Any geographical dataset provides information about one or more characteristics f at every location (x,y) within the footprint; theme defines the nature of f, or the dataset's *semantics*. Geographical themes range from land surface elevation to soil class, land cover class, or population density. The specification of geographical theme is complicated, however, by three issues.

First, geographical themes lack a controlled vocabulary that is comparable to those of library subject classification. There are no accepted standards for themes, and the wide range of possible themes makes it very difficult to develop one. Second, there is a tendency for GIBOs to provide information on more than one theme, either through the lumping of many *layers* into a single database, or through the assignment of multiple attributes to a single set of objects. This issue might be dealt with by changing the *granularity* of data, by breaking up a database into constituent layers and thus better-defined themes. But it is not clear that it would result in a better search process. Finally, there is the possibility that the user will define a new theme by interpretation or manipulation of the raw data, a practice that is common, for example, in the use of remotely sensed imagery or aerial photography (Curran et al, this volume). Thus a dataset that is classified as 'aerial photograph' might be used to provide information on land cover type if the

user were willing to attempt an appropriate classification. In principle, therefore, every dataset should be classified by all of those themes that can be derived from the data by processing.

Theme is a discrete, nominal variable, and the goodness of fit of a GIBO's theme to the user's specification can only be measured in a binary fashion. Date is interval, however. Thus the user might specify a range of acceptable dates, and the GIBO would be identified as a 'hit' if its date fell within the range.

Any geographical dataset must have an associated *level of detail*, S. The concept of *scale* was introduced earlier with respect to the user's ability to model process, and with the implication that any dataset with a level of detail equal to or finer than P would be acceptable as input. Goodchild and Proctor (1997) discuss the measurement of level of detail in digital geographical datasets, and conclude that a linear measure is most appropriate, despite the massive legacy of representative fraction as a characteristic of paper maps. They also discuss the difficulties of measuring S for irregular geographical data models. An ideal fit to the specification would have $S=P$; in practice, however, the requirement is unlikely to be met perfectly, and instead the suitability of a GIBO with level of detail S against a requirement for data at scale P is some decreasing function $g(P-S)$ for $P>S$ (greater implies coarser or more generalised when detail and scale are expressed as linear measures). When $P<S$ the dataset may still be useful if the user has access to techniques for simulating the missing detail, and thus measuring the impact of the lack of sufficiently detailed data (Ehlschlaeger et al 1997).

In summary, the need for a geographical dataset can be specified in terms of footprint, theme, date and level of detail, plus other more specialised elements as appropriate (an exhaustive list is provided by the US Federal Geographic Data Committee's (FGDC) Content Standards for Digital Geospatial Metadata, *http://www.fgdc.gov*); and candidate information objects can be specified through *metadata* that are defined in the same terms. Goodness of fit can be measured as a binary property in some cases; as a function in the case of level of detail; and as a normalised ratio of intersection area in the case of footprints. Since there is almost no likelihood that a GIBO will match perfectly to an independently specified requirement, but a substantial likelihood that more than one dataset will be identified as potentially useful, some means must be devised for weighting these components of goodness of fit for alternative candidates, ranking the totals, and making a rational choice. Let x_{ij} denote the result of comparing $GIBO_i$ to the specification on the jth metadata component. Then the goodness of fit G_i will be a function $G(x_{i1},x_{i2},\ldots)$.

4.3 Search

Armed with tools for specifying need and measuring the goodness of fit of candidate GIBOs, the discussion now turns to the process of search. There are now thousands of sites on the WWW offering GIBOs, some with restrictions on use, and some charging for use, but many offering data at no cost and without restriction. The US National Geospatial Data Clearinghouse (NGDC; *http://www.ngdc.gov/clearinghouse/clearinghouse.html*) is one example, and many others exist in other countries, at other levels of government, and in other agencies.

Projects such as NGDC are based on the principle of *one-stop shopping*, that is, the principle that a user within some geographical domain and in need of geospatial data would go to a single, known source. All available data covering the domain would be catalogued by the source, and there would be an implicit guarantee that if the data could not be found in that source, they could not be found anywhere.

Unfortunately the one-stop shopping model is likely to fail, for several reasons. First, there is no rational basis for assigning this function to any one level in the administrative hierarchy. While it might make sense for datasets covering an entire nation to be accessible through a national server, by the same principle datasets covering an entire county should be accessible through a county server, not a national server; and what about datasets covering parts of several administrative units? The connectivity of the Internet is not perfect, and users in a given county would clearly not welcome being asked to store and access all data in one massive, global server, or the loss of control and custodianship that this would imply. Goodchild (1997) argues that the rational solution to this problem assigns each GIBO to a single server somewhere within the GIBO's footprint.

Second, projects such as NGDC are designed to serve only geographical data. While it was argued earlier that the needs of geocomputation are likely to be almost exclusively for geographical data, it does not follow that it is optimal to serve such data from exclusive servers. Goodchild (1998a) has argued that mechanisms devised for searching for geographical data can also be used to search for other types of information that are not geographical, but that nevertheless possess geographical footprints; he terms these *geographically referenced* datasets.

Finally, such projects require a high level of conformity among those who make use of them to serve data. There is a large expense in building specifications for GIBOs, particularly specifications with the richness of the FGDC metadata standard. Few incentives exist to create these specifications, other than the knowledge that by doing so one makes one's data more accessible to others. Given a choice, the custodian of data may elect to mount the data only on a small, personal server, and to provide only minimal documentation, letting the potential user bear any of the risks associated with use.

In short, any search for specified geographical data is likely to have to consider the possibility that a suitable GIBO may exist on any one of a large number of possible servers. Some means for directing the search is therefore necessary. The next sections consider two possible alternatives.

4.3.1 Search Engines

One of the most useful ways of searching for information on the WWW is to access a search engine, one of a number of sites that offer directories to the WWW. Current search engines are able to catalogue the WWW's contents by sending out intelligent agents, or *web crawlers*, to find and abstract the information available at WWW sites. They do this by following hyperlinks, or links that the custodians of

sites have put in place to link to information at other sites. Information that is not linked is in a sense invisible, since web crawlers will not find it.

Web crawlers assume that information at WWW sites is in the form of text, and attempt to identify key words and phrases. The user of a search engine specifies a word or phrase, and the engine returns a list of sites determined to have information in which that word or phrase appears, in what is determined to be a significant manner (e.g. in the title of a page). Certain words and word forms are clearly more useful for this kind of search; a person's name, or a number, may be much more useful than a common word.

The catalogues produced by these search engines are very different from the metadata specifications discussed in the previous section, or the catalogues of the traditional library. There is no separation between footprint, theme, date and level of detail; instead, words and phrases must serve all purposes, and there is no guarantee that a word extracted from a body of text will in any way characterise the entire text. Moreover, GIBOs are built using the data models of GIS; if text exists in a GIBO, it does so in a very limited way in the form of attributes, or in metadata. Search engines have not been designed to abstract useful metadata from GIBOs.

Several authors have commented on the potential for a new generation of search engines that could seek out and catalogue GIBOs, generating something much closer to a metadata specification. The web crawlers associated with such a search engine would have to be able to recognise GIBOs, and to open them in order to define the key descriptors of their contents. This would clearly be much easier to do if the custodian had provided metadata in some standard format; and much easier for some geographical information formats than others. For example, a web crawler that can open a GIBO containing coordinates in some standard Earth system, such as latitude/longitude, can determine the bounding coordinates of the GIBO; but this is clearly not possible for a raster image that has no tie to the Earth's surface.

4.3.2 Collection-Level Metadata

The user of a research library has certain informal expectations about the information it is likely to contain. There is an assumption, for example, that a library at a research university will contain all important journals, and all significant books; the degree to which it does so is a commonly used measure of its success. A library will also have special collections, many of which will be unique; their existence will be known to researchers in the appropriate subject areas.

This heuristic breaks down almost completely in the digital world of the WWW. Since all users can in principle be served from a single site on the Internet, there is no need for sites to duplicate each other's contents. Instead, all WWW sites are to some degree analogous to the library's special collection, but their sheer numbers make the task of knowing which site has what virtually impossible.

In the case of GIBOs the geographical nature of the information may provide the basis for an effective heuristic that can be used to limit search. Goodchild (1997) has defined *information of geographically determined interest* (IGDI) as an

information object that is of greatest interest to users in the immediate vicinity of its geographical footprint; GIBOs are a subclass of IGDIs. The servers most likely to contain a given GIBO are those closest to its footprint; and the size of the footprint also provides an indication of the level in the administrative hierarchy that is most likely to serve the GIBO. Unfortunately the architecture of the Internet makes it impossible to direct search by geographical location, although some Internet domains are geographically defined. But research efforts are currently underway to develop appropriate protocols (Navas and Imielinski 1997) that would make this much more feasible.

4.4 Fitness for Use

The assessment of fitness for use has been modelled as a comparison between the user's specification, as expressed in metadata, and the specifications of candidate datasets, with an associated metric. While some components of metadata imply a binary assessment (e.g. date), the results of other comparisons must be measured on continuous scales, leading to a ranking of candidates. Systems such as the Alexandria Digital Library (*http://alexandria.ucsb.edu*) return a number of potential candidates, ranked by a measure over which the user has some control, and limited by parameters, such as the maximum number of candidates, that are also controlled by the user.

The process of search in a library is inherently 'fuzzy' or uncertain, and it is frequently necessary for the user to *browse* in order to locate a suitable book or article. In effect, the user is unable to make a complete specification in advance, and instead refines the specification during the search process. Libraries support browsing by arranging to shelve books on similar subjects together, so that the effort on the user's part in accessing several books on the same subject is not much greater than the effort in accessing one. Similarly, a search in a digital domain should return several GIBOs, each with high score G, and allow the user to open them, browse their contents, and possibly refine the search criteria as a result. The issue of browse is discussed again later in the context of retrieval.

While any GIBO can be assessed against a specification using the methods discussed earlier, the user will also need some assurance that the GIBO's metadata are complete and accurate. It is likely that the metadata will not be complete in many cases, either because the custodian did not provide a complete specification, or because a search engine was unable to determine one. It is also possible that the data do not meet the claims made in the metadata, because of high levels of error, or because the metadata are simply incorrect.

Goodchild (1998b) has reviewed the description of data quality in metadata, and the issues involved. Goodchild et al (1998b) note that the literature on geographical data quality now includes many models, with many associated parameters, and it is increasingly unlikely that a user of geographical data would be sufficiently knowledgeable in this area to make effective use of full data quality information. Instead, they argue that data quality might be described by a *process* rather than a set of parameters. The process would be encapsulated with the data, in Java or some

other code that can be executed on any client; and by initiating the process, the user would generate a number of realisations of an error model defined by the custodian or producer of the data. They argue that such a process is a full and complete specification of data quality, and yet requires no expert knowledge on the part of the user. Following Openshaw (1989), they argue that error model realisation provides a comprehensive approach to the data quality problem.

4.5 Retrieval

GIBOs have a tendency to be large; a complete representation of the US street network occupies some 10^{10} bytes, for example, and a complete Landsat scene some 300 Mb. The bandwidth available between server and client may be constrained by a modem, and will be subject to contention from other users. Thus delivery of a selected GIBO is often far from a trivial technical issue. In addition, the user may not be certain that the GIBO meets requirements until it is opened, since there will always be ambiguity and perhaps inaccuracy associated with metadata descriptions.

Server-side processing may address many of these issues, allowing the user to specify a window or other basis for selection of part of a GIBO. This is a service for which there is no obvious analogue in the traditional library, since one can only deliver part of a physical object by destroying the object, although there are many examples of server-side processing in data dissemination. It would be simple, for example, for the server to send only that subset of B_i contained in A. The degree of overlap between B_i and A has already been used in the proposed measure of 'goodness of fit'; unfortunately, cases where there are large economies to be gained by clipping B_i to A are also cases where B_i will have been given a low score because of low overlap.

More comprehensive are methods of *progressive transmission*. Suppose B_i could be organised in a hierarchical fashion, beginning with a representation of the coarsest spatial variation, and progressing to the finest details. The coarsest components would also be comparatively small in volume, and could be transmitted quickly. The user could be given the option of stopping the transmission at any point, if the GIBO appeared to be inappropriate for the requirement. Virtually any hierarchical decomposition will meet the needs of progressive transmission, including quadtrees (Samet 1990) and wavelets (Chui 1992). Unfortunately, while many suitable techniques exist for raster data, the problem of hierarchical decomposition and progressive transmission of vector data seems much more difficult, since it is essentially the problem of automated cartographic generalisation (Müller et al 1995). The viewpoint-centred methods often used in visualisation, which generalise the periphery of the field of view and thus reduce data volume, are inappropriate for geocomputation, which almost certainly requires uniform coverage of the study area.

Progressive transmission can help in two ways, by providing coarse approximations to a GIBO that the user can examine and assess, and by allowing the user to truncate transmission when some acceptable level of detail has been reached. Wavelets and other efficient decompositions have no overhead, since the hierarchically structured data occupies the same volume as the conventional form. The concept will be familiar to WWW users, since it is embedded in some image transmission standards.

4.6 Opening

Opening is the final step in the five-stage process of data acquisition for geocomputation. Having retrieved a GIBO, the user is concerned with the task of opening it in some local application, for visualisation, statistical analysis, or input to a simulation model. Note that the ability to open the GIBO has already been assumed in the previous section, where the user was able to make an informed decision based on the GIBO's contents.

It is possible to define various levels of opening. For example, the user may be able to display the form of the dataset by interpreting the coordinates defining its basic objects, but unable to interpret the attributes because no information is available defining the GIBO's semantics. Many of the details needed for successful opening, such as the name of the GIS used to create the data, may be contained in the GIBO's metadata, and transmitted as part of the *wrapper*. The Open GIS Consortium (OGC; *http://www.opengis.org*) is actively developing the standards that will allow a user application to open GIBOs of a wide range of formats and origins without any intervention on the part of the user, but it will be some time before this transparency becomes part of standard practice in geocomputation.

Opening also has no analogue among the services of the traditional library, whose responsibilities normally end when the information object is put into the hands of the user. In a digital world it is possible to imagine a host of client-side and server-side services concerned with processing information derived from distributed stores. But this suggests an awkward problem noted earlier: to what extent should metadata define not only the attributes of the GIBO, but also any form into which the GIBO can potentially be restructured or manipulated? As Kuhn (1997) argues, two datasets are essentially identical if one can be manipulated to provide the same information as the other, no matter what their actual structures may be, provided the costs of manipulation are not considered.

4.7 Conclusion

This chapter has reviewed the issues raised by a massive shift in the arrangements for retrieval of geographical information as input to geocomputation. Traditionally, the services provided by a library allowed its users to retrieve certain types of information, on the understanding that that information was to be found within physical volumes. Because digital data presented its own peculiar difficulties, early approaches to data dissemination emphasised large data centres and archives, with their own largely unique protocols. The library model did not work well for geospatial data, because of the problems of effective handling and cataloguing of

maps and images, and so early efforts to adopt digital technology in support of dissemination of geospatial data occurred mostly outside the library domain.

In the past few years the growth of the WWW has opened the possibility of a generic approach to information dissemination, in which the nature of the information is largely irrelevant to the process of retrieving it. Information objects consist of 'bags of bits', with wrappers that define important characteristics of the bag's contents to the digital environment. Geographical information is thus in principle as easy to find, assess, retrieve and use as any other kind of information once it is in digital form, and many of the old distinctions based on analogue media are being questioned, reassessed, or ignored.

A five-stage process has been presented, and the chapter has discussed issues that arise at each stage, based largely on the author's experience with the Alexandria Digital Library project. While it is possible to see how each of the five stages might operate, and many efforts are under way to facilitate many of its elements, it will be many years before the legacies of previous arrangements, the problems with competing and incompatible standards, and resistance to the effort involved in providing the necessary metadata disappear, if they ever do. Until that happens, the process of search for data to support geocomputation will continue to rely, as it always has, on networks of personal contacts, idiosyncratic knowledge, luck, and the expertise of SAPs.

References

Chui C K 1992 *An introduction to wavelets*. Boston, Academic Press

Delcourt H R, Delcourt P A, Webb T III 1983 Dynamic plant ecology: the spectrum of vegetation change in space and time. *Quaternary Science Review* 1: 153

Ehleringer J R, Field C B (eds) 1993 *Scaling physiological processes: leaf to globe*. San Diego, Academic Press

Ehlschlaeger C R, Shortridge A M, Goodchild M F 1997 Visualizing spatial data uncertainty using animation. *Computers and Geosciences* 23: 387–95

George P L 1991 *Automatic mesh generation*. New York, John Wiley

Goodchild M F 1992 Geographic data modeling. *Computers and Geosciences* 18: 401–8

Goodchild M F 1997 Towards a geography of geographic information in a digital world. *Computers, Environment and Urban Systems* 21: 377–91

Goodchild M F 1998a The geolibrary. In Carver S (ed.) *Innovations in GIS 5*. London, Taylor & Francis 59–68

Goodchild M F 1998b Communicating the results of accuracy assessment: metadata, digital libraries, and assessing fitness for use. In Congalton R G, Mowrer T (eds) *Spatial accuracy assessment in natural resource analysis*. Chelsea (US), Ann Arbor Press (in press)

Goodchild M F, Proctor J D 1997 Scale in a digital geographic world. *Geographic and Environmental Modelling* 1: 5–23

Goodchild M F, Montello D R, Fohl P, Gottsegen J 1998a Fuzzy spatial queries in digital spatial data libraries. *Proceedings, IEEE–FUZZ, Anchorage, May 4–8*

Goodchild M F, Shortridge A, Fohl P 1998b Encapsulating simulation models with geospatial datasets. *Proceedings, Third International Symposium on Spatial Accuracy Assessment in Natural Resources and Environmental Sciences*

Knupp P, Steinberg S 1993 *Fundamentals of grid generation*. Boca Raton, CRC Press

Kuhn W 1997 Approaching the issue of information loss in geographic data transfers. *Geographical Systems* 4: 261–76

Müller J C, Lagrange J P, Weibel R (eds) 1995 *GIS and generalization: methodology and practice*. London, Taylor & Francis

NRC (National Research Council) 1995 *A data foundation for the national spatial data infrastructure*. Washington, DC, National Academy Press

Navas J, Imielinski T 1997 GeoCast – Geographic addressing and routing. *Proceedings, MOBICOM 97, Budapest, Hungary*: 66–76

Onsrud H J, Rushton G (eds) 1995 *Sharing geographic information*. New Brunswick, NJ, Center for Urban Policy Research

Openshaw S 1989 Learning to live with errors in spatial databases. In Goodchild M F, Gopal S (eds) *Accuracy of spatial databases*. London, Taylor & Francis: 263–76

Raper J 1997 Progress towards spatial multimedia. In Craglia M, Couclelis H (eds) *Geographic information research: bridging the Atlantic*. London, Taylor & Francis: 525–43

Rosswall T, Woodmansee G, Risser P G (eds) 1988 *Scales and global change*. New York, John Wiley

Samet H 1990 *The design and analysis of spatial data structures*. Reading (US), Addison-Wesley

Smith T R, Andresen D, Carver L, Dolin R et al 1996 A digital library for geographically referenced materials. *Computer* 29(7): 14