CHAPTER FIVE

The geolibrary

MICHAEL F. GOODCHILD

5.1 INTRODUCTION

While a geographic data set is defined as a representation of some part of the Earth's surface, many other types of information also refer to specific places on the Earth's surface, and yet are not normally included in discussions of geographic databases. They include reports about the environmental status of regions, photographs of landscapes, guidebooks to major cities, municipal plans, and even sounds and pieces of music. All of these are examples of information that is geographically referenced, or georeferenced for short, because it has some form of geographic footprint. Clearly geographic information is by definition a subset of georeferenced information.

Footprints can be precise, when they refer to areas with precise boundaries, or they can be fuzzy, when the limits of the area are unclear. For example, a municipal plan for a city has a precise footprint, but tourist information on 'Northern California' does not, because 'Northern California' is not precisely defined on the ground. Data sets can have more than one footprint, or a footprint that is more complex than a single polygon. For example, the footprint of Handel's Messiah might include the place where it was composed (London), as well as the place where it was first performed (Dublin), and perhaps also geographic references to the life of its composer.

Georeferenced information is likely to be more pertinent in some places than others. Generally, interest depends on how far one is from the footprint, and also on the size of the footprint. In California, interest in a guidebook to France will be much greater than in a municipal plan of the city of Rouen. A bookstore or public library in California might carry such a guidebook, but almost certainly no bookstore or public library anywhere in the state would have the municipal plan of Rouen. In this sense there is a sharp distinction between georeferenced data sets and other types of information: a mathematical encyclopaedia, for example, is presumably of uniform interest everywhere on the planet.

A geolibrary is a library filled with georeferenced information. Information is found and retrieved by matching the area for which information is needed with the footprints of items in the library, and by matching other requirements, although the footprints always provide the primary basis of search. A geolibrary can handle queries like 'what information do you have about this neighbourhood?', 'do you have a guidebook covering this area?', 'can I find any further information about the area in which the Brontë sisters

lived?', or 'what photographs do you have of this area?' In all of these queries the geographic footprint provides the primary basis of search.

Recent years have brought an explosion of available information, notably in the form of information accessible through the World Wide Web (the 'Web'). But while certain tools exist for browsing and searching this immense information resource, more generally there is an acute shortage of methods for organizing, managing, cataloguing, and integrating data. The traditional library has perfected the cataloguing of books into a fine art, based on author, title, and concepts of subject. But the current generation of search engines, exemplified by Alta Vista or Yahoo, are limited to the detection and indexing of key words in text. By offering a new paradigm for search, based on geographic location, the geolibrary might provide a powerful new way of retrieving information.

Consider an example of crisis management. The Oklahoma City bombing of April 1996 created an immediate need for information about the construction of the building, nearby hospital resources, evacuation routes, the locations of underground gas pipes that might have been damaged, the potential for adverse weather to affect the rescue effort, the identities of employees, and much more. The common element in all of this information was the geographic location of the explosion – if this could have formed the basis of search, the relevant information might have been assembled much more rapidly.

The purpose of this paper is to explore the concept of the geolibrary, and its relationship to the traditional concerns of GIS research, as a thought-experiment. Many elements of the geolibrary are present in traditional institutions, data centres, and earlier research efforts, and these are reviewed in the next section of the paper, which explores the degree to which geolibraries can be said to exist. This is followed by a further discussion of the value and impact of a functioning geolibrary. The subsequent section identifies the components of the geolibrary, and this is followed by a discussion of the important research issues that must be solved if the geolibrary is to be achievable.

5.2 DO GEOLIBRARIES EXIST?

A user of a geolibrary might begin the process of search and retrieval by pointing to a map or a globe. Physically, this would mean having a large map or globe in the library's entrance; but it would be impossible to link the action of pointing to a map or globe with the card records of appropriate items in the library. It would be impossible also to provide for many different levels of detail, in order for the library user to be able to specify areas as large as France or as small as a single city neighbourhood.

An alternative approach would be for the user to specify a placename, and to provide a fourth kind of indexed card catalogue, in addition to the conventional indexing by author, title, and subject. But placenames only work to a degree. There would be no way, for example, to link entries under France with entries for the various French Departments, or the names of European regions that include part or all of France. Geographic placenames do not lend themselves readily to use as subject headings in card catalogues, because of the complex set of hierarchical and other relationships that exist between geographic locations.

Essentially, it is impossible to build a physical geolibrary, although conventional map libraries come as close as it is possible to come. In a digital world, however, these problems disappear. The user of a digital geolibrary can be presented with a globe; can zoom to the appropriate level of detail; can access lists of placenames and see their footprints; and can move up or down the placename hierarchy using links between places.

Moreover, a digital geolibrary solves the problem of physical access, if the services of the library are provided over a universal network like the Internet. In a digital world, the differences between storage media, which have made it difficult for the conventional physical library to handle music, maps, or photographs, disappear. And, finally, the collection of a geolibrary can be dispersed – a digital geolibrary can consist of a collection of servers, each specializing in materials about its local region.

5.3 THE ALEXANDRIA DIGITAL LIBRARY

In the past few years there has been an increasing awareness of the technical feasibility of digital libraries – that is, libraries that emulate the services of conventional physical libraries, but using digital technology, with access through digital communication networks. Such digital libraries would be of immense benefit in extending access to library services to anyone connected to digital networks; would add the capability to search information by content; and would provide information in digital form, allowing users to analyze it further, using suitable application software.

The Internet has already moved us closer to the vision of universal access via electronic networks (although many would argue that it has simultaneously widened the gap between have's and have-nots). A simple calculation shows that the textual contents of a typical, large research library, when digitized in a simple code such as ASCII, amount to order 10^{13} or 10^{14} bytes, and thus lie within today's storage capacities. In short, a digital library is sufficiently feasible to justify further research. In the USA in 1994, three federal agencies (the National Science Foundation, NASA, and the Advanced Research Projects Agency) jointly initiated a Digital Library Initiative, funding six projects for a period of four years. The Alexandria Digital Library (ADL), based at the University of California, Santa Barbara, is one of those projects.

The primary goal of ADL is to develop the services of a distributed library of georeferenced materials. Smith *et al.* (1996) describe ADL, and further information and descriptive reports, as well as access to the current prototype, are available via http://alexandria.ucsb.edu.

In the first phase of the ADL project, from October 1994 to March 1995, we developed a Rapid Prototype using existing off-the-shelf software, including ArcView 2, Tcl/Tk, and Sybase. This was a stand-alone prototype, distributed on CD for Windows. Our first Web prototype was completed in 1996, using a standard Web browser but providing full Internet access. ADL is currently (April 1997) working on a new Web prototype using Java and VRML tools.

At this point in its development, ADL satisfies some of the requirements of a geolibrary, but by no means all. A subsequent section, which reviews the necessary components of a geolibrary, includes an evaluation of the relevant components of ADL in its current prototype.

5.4 OTHER PROJECTS

ADL is by no means unique in its attempt to provide the services of a geolibrary. The DLI project at the University of California, Berkeley, also focuses on multimedia and georeferenced materials, although it has additional emphases (information on all of the DLI projects can be accessed through the ADL web site). Many developers of software for processing maps, images, and similar materials have recognized the need for additional

modules to manage what is often a large and rapidly growing resource of data sets. ESRI, for example, offered MapLibrarian as a component of early versions of its popular ARC/INFO software, and Intergraph provides its own document management software to meet the needs of its larger CAD and GIS customers. Elements of the geolibrary concept are present in many systems for management, browse, and dissemination of geographic data, including EOSDIS (the Earth Observing System Data and Information System), GC-DIS (the Global Change Data and Information System), the FGDC's (Federal Geographic Data Committee's) National Geospatial Data Clearinghouse, and many others. One of the first such projects was MARCMAP (Morris *et al.*, 1986), at the University of Edinburgh beginning in the early 1980s (and see http://www.geo.ed.ac.uk/~dcf/gisa/Cartonet.html); and see also Goodchild and Donkin (1967).

5.5 THE VALUE OF A GEOLIBRARY

In the first instance, a geolibrary could provide an interesting new access mechanism to the contents of a conventional library – for example, a tourist planning to visit Paris would have a powerful new way of finding guidebooks. But the full power of the geolibrary lies in its ability to provide access to types of information not normally found in libraries.

First, the geolibrary would be a multimedia store. Because everything in a digital library is stored using the same digital medium, there are none of the physical problems that traditional libraries have had to face in handling special items like photographs, sound, or video.

Secondly, the geolibrary could focus on building a collection of items of special or local interest. In a networked world there is no need for libraries to duplicate each other's contents – it may be sufficient for one digital copy of the Rouen municipal plan to be available in a digital geolibrary located in Rouen, or perhaps in Paris, provided other geolibraries know it is there. When material is georeferenced, there is a clear advantage to locating its digital version on a server close to, or within, the geographic footprint of the material, since that is where interest in it is likely to be highest.

Thirdly, because interest in it tends to be geographically defined, much georeferenced information is unpublished or fugitive. The geolibrary provides an effective mechanism for collecting such information in special, local collections, and making it widely available.

In summary, the contents of a geolibrary would be very different from those of a conventional physical library. They would be dominated by multimedia information of local interest – in fact precisely the kinds of information needed by an informed citizenry, one that is deeply involved in issues affecting its neighbourhood, region, and planet. Because its contents would be different, a geolibrary might attract an entirely new type of library user. At the same time, geolibrary technology would be of great benefit to major utilities, resource industries, and other private sector corporations with large information management problems and a strong geographic context. Such corporations, running geolibraries over intranets, might provide the market needed to stimulate commercial development of geolibrary technology.

5.6 THE COMPONENTS OF A GEOLIBRARY

While there are sharp differences in approach and scope between the various projects that have attempted to construct versions of a geolibrary, there is now among them a degree of consensus on its basic nature.

5.6.1 The browser

Browsers are specialized software applications running locally in the user's computer, and providing access to the Web. The Alexandria Digital Library project's current prototype is an example of a geolibrary using standard Web browser software, such as Microsoft's Internet Explorer. Ideally, a geolibrary browser would be more specialized, but universally available. It would have 3D extensions (e.g. a VRML plug-in) to allow the user to interact with the surface of the globe, and for specialized interactions such as 'fly-bys'. To avoid having to communicate large amounts of repetitive data, it might use a 'hybrid' approach by storing the basemap and gazetteer locally (see below). Finally, like current Web browsers it would provide a uniform 'look and feel' across a wide range of computer hardware.

5.6.2 The basemap

The basemap provides the image of the Earth on which the browser's user can specify areas of interest, and on which footprints of library items can be displayed. Its level of geographic detail defines the most localized search possible. It should include all the features likely to be relevant to a user wanting to find and define a search area, including major topographic details and placenames. The importance of such features will vary between users, as will levels of detail, so it will be necessary to establish protocols that allow use of specialized basemaps for particular purposes.

Basemap information may be voluminous and is not likely to change frequently, so rather than transmit it repeatedly from a server it may be more efficient to store it locally, in a specialized 'hybrid' browser.

Suitable basemaps include digital topographic maps, and also images of the Earth's surface. Additional detail can be provided by digital elevation data, so that the basemap provides a close resemblance to the actual surface of the Earth. The availability of such data is discussed below.

5.6.3 The gazetteer

'Gazetteer' is a technical term for the index that links placenames to a map. Gazetteers are commonly found in the backs of published atlases and on city maps. In a geolibrary the gazetteer allows a user to define a search area using a placename, instead of by finding the area on the basemap. The gazetteer may include placenames that are not well defined.

For use in a geolibrary a gazetteer must include extents, or digital representations of each placename's physical boundary. Links between placenames allow searches to be expanded or narrowed – they can be vertical, identifying places that include or are included by other places, and also horizontal, identifying neighbouring places.

A gazetteer protocol would allow specialized gazetteers to be used in particular applications. Like basemaps, gazetteers are not likely to change rapidly, so could be stored locally in 'hybrid' browsers. The availability of suitable data is discussed below.

5.6.4 Local collections

A geolibrary could operate in a standalone form, with the browser providing access to a collection of items of information stored on the same machine, or on a machine linked

logically to it. The browser would display the availability of information in the form of icons superimposed on the basemap, with associated details, and the user would be able to retrieve and manipulate the items.

5.6.5 Server catalogues

The contents of a specialized geolibrary collection could also be made available to browsers over a network. Each collection would be maintained by a server. A user would initiate a search with a browser, and transmit a specification of the search to a server, which would then respond with an indication of the items in its collection satisfying the search. These would then appear as symbolic icons superimposed on the browser's basemap, and as items in a list with additional details. The user could request more detail, including the downloading of the item itself, by selecting an icon from the basemap or from the list. In addition, a server might provide a generalized 'browse' version of the item to avoid forcing the user to wait for a lengthy download (see below).

For this mechanism to work, the user must give search specifications in a standard format, following an agreed protocol. The basis of such protocols already exists in standards such as MARC and the Federal Geographic Data Committee's Content Standards for Digital Geospatial Metadata (http://www.fgdc.gov), and in projects such as ADL.

5.6.6 Basemap data

Several suitable sources of data exist for basemaps:

- *Digital topographic data.* Available for the entire land area of the planet at 1:1 000 000 in the Digital Chart of the World (in very rough terms, a map at a given scale shows features that are at least 0.5 mm across at the scale of the map; this means that a 1:1 000 000 map would show features at least 500 m across). Data are also available for smaller areas at larger scales – e.g. for the continental U.S. at 1:100 000, or for the UK at 1:10 000.
- *Imagery.* Available from the Landsat satellite at 30 m resolution; from the SPOT satellite at 10 m resolution; from Russian satellites; and in late 1997 for selected areas at 1 m resolution.
- *Digital elevation data.* Available for parts of the USA at 30 m resolution; for the entire planet (classified) at approximately 100 m resolution; for the entire planet (public domain) at 10 km resolution; expected to be available in late 1997 for the entire planet at 30 m resolution.

5.6.7 Gazetteer data

A combination of public sources (the Geographic Names Information System of the USGS, and the Board on Geographic Names) can provide about 7 million point records. There are no significant sources of information on feature extents. There are significant problems to be overcome in dealing with varied alphabets and languages.

5.6.8 Search across geolibraries

On the Web one has access to numerous search engines to assist in the process of identifying the server most likely to contain a given set of information. Search engines such as AltaVista work by 'crawling' the Web looking for servers, and extracting appropriate key words that can be indexed. This same mechanism would work for placenames, but would be subject to the same problems identified earlier that limit the ability of a placename index to find georeferenced information successfully.

Instead, a geolibrary must use alternative mechanisms to identify the servers to be searched. One would be the creation of a union catalogue – the US National Geospatial Data Clearinghouse is a current example – that contains a record for each item available in any server whose contents have been registered with it.

A second mechanism would rely on the development of a set of rules for predicting the server(s) most likely to contain a given item. Such rules could depend on geographic location – the server most likely to contain an item about location X would be the one closest to X, for example. This approach will work only if the set of servers is small, and well organized.

Thirdly, the search could visit every known geolibrary for every search. This is clearly time-consuming, and fails to 'scale' as the number of geolibraries increases.

Fourthly, one could introduce extensions to the mechanisms currently used by the Web crawler agents acting for the search engines. These would include explicit identification of placenames, and mechanisms for identifying the co-ordinates of footprints embedded in text. This option seems the most powerful, but would require development of a new generation of crawlers and associated protocols.

5.6.9 Browse images

The bandwidth available between a server and a client varies enormously, depending on many factors. Geographic information can be voluminous, and it is easy for a single multiband image of a small part of the Earth's surface to exceed a gigabyte. Thus, a geolibrary operating over an open network will have to consider the problems of limited bandwidth in its design. At the very least, a user should be warned about the probable time required to download an information object.

One general solution to these issues is to support progressive transmission. In such schemes, a coarse and thus comparatively small version of the information is sent first, followed by increasing levels of detail. In the case of an image, the technology of wavelet decomposition provides a suitable mechanism. In ADL, we also use a concept of a browse image or thumbnail sketch, a small raster containing a very generalized version of the image, and make this available to the user who expresses interest in a particular information object. There are problems, however, in generalizing the idea to types of information other than images.

5.6.10 Objects and wrappers

In the conventional library, information is packaged in convenient bundles as bound volumes. In the geolibrary, each information object, or 'bag of bits', must be packaged inside an appropriate digital wrapper, which provides the information needed by the

system to ensure appropriate handling. The wrapper must identify, for example, the internal format of the data in sufficient detail to allow the system to open and display its contents. ADL currently supports only a few formats. On the other hand, it seems very unlikely, given the philosophy of the geolibrary and other Web-based systems, that one would ever be able to impose or achieve uniform format standards. In the long run, therefore, a geolibrary must be able to support a large number of alternative information object formats.

5.7 RESEARCH ISSUES

Besides the technical problems identified above, a large number of research questions need to be addressed.

1 What are the legal, ethical, and political issues involved in creating geolibraries? What problems must be addressed in the area of intellectual property rights, and what are their technical implications?

2 What are the economics of geolibraries? Who should pay for their creation and maintenance? What elements might be free (funded by the public sector, or by the private sector as loss-leaders)? Where are the potential revenue sources?

3 What is the appropriate scale for a geolibrary? How much material should be assembled on one server, and what geographic area should it cover? What are possible transition trajectories between current arrangements and a future distributed geolibrary?

4 What institutional structures would be needed by a geolibrary? What organizations might take a lead in its development? What is a possible timetable?

5 What models exist for the creation of geolibrary metadata? What are the advantages of following a library cataloguing model, versus a fully voluntary model where metadata is provided by the custodian? How much metadata is needed?

6 What techniques can be used to locate the geolibrary most likely to contain a certain item (see 5.6.8 above)? Research would include experiments with working prototypes of the most promising approaches.

7 What are the cognitive problems associated with using geolibraries? Is it possible to construct a geolibrary that is useful to a child in Grade 3, for example? What protocols would users have to master, and what problems would occur in using geolibraries across cultural or linguistic barriers?

8 Is it possible to compare the geolibrary with conventional libraries, both physical and digital? What are appropriate metrics of comparison, and what general principles can be learned from their application?

9 What protocols are needed to support the geolibrary? What role should existing institutions play in developing them? Are existing protocols such as HTTP and VRML sufficient? What protocols are needed for metadata, base maps, and gazetteers?

10 Do the data sets necessary to support the geolibrary exist? What initiatives are needed to develop or compile them?

11 Is it possible to develop a new generation of search engines and agents that can scan the Web for georeferenced materials? What protocols would be needed to make such agents effective?

12 Can the concept of a footprint be extended to the indeterminate or fuzzy case? What mechanisms are needed to define, store, and use fuzzy footprints?

13 Is a Boolean approach to the intersection of query and information object footprints appropriate, or is it possible to develop metrics of fit between them, and to rank information objects on that basis?

14 Can the concepts of browse images and progressive transmission be extended from the raster case, notably to vector data sets, and non-geographic information objects? What other options exist for intelligent use of bandwidth in geolibrary applications?

5.8 CONCLUSION

Projects like ADL, and the concept of a geolibrary, are bridges between the GIS and library communities. Libraries have centuries of expertise in such functions as quality assurance in building collections, and abstracting of information to support search, that are as relevant for geographic information as they are for any other type. Yet, as in so many other areas, traditional libraries have been constrained by physical limitations to focus on certain types of information that are comparatively easy to handle. The digital revolution has changed all that, and has made geographic information potentially as accessible and easy to handle as other types. At the same time digital communication networks have provided the potential for universal access to information. Thus, we find ourselves in the GIS research community at the beginning of a period of exciting collaboration with the library and information science communities.

At this point in history, libraries are faced with apparently insurmountable problems (Hawkins, 1996). The published corpus of humanity is growing rapidly, and doubling in not much more than ten years. Journal prices continue to rise at well above the rate of inflation. Library budgets are contracting, and libraries are faced with unprecedented problems of security. The pressures to find new approaches, and to take advantage of new technologies, are high.

I would like to close with four points. First, the concept of georeferenced information extends the domain of digital geographic information well beyond more conventional digital maps and images, to include a vast array of information with definable geographic footprints. This new class offers an exciting new domain for the GIS research community.

Secondly, the geographic key is a powerful way to organize and integrate information, whether it be geographic or georeferenced. It has been difficult to exploit this particular key in the past because it is based in a continuum, whereas the more successful keys of information management – author, title, subject – are inherently discrete. But in a digital environment this problem is much less severe. Moreover, the Web has shown how important it is to develop new ways of integrating information, to support effective search over a rapidly expanding but inherently unorganized information base.

Thirdly, geolibrary projects like ADL demonstrate the potential for radically different approaches to information management. Information access mechanisms ultimately determine the kinds of material that can be served by systems such as libraries – new mechanisms thus offer the potential for storing novel types of information, and for reaching new user communities.

Finally, I hope this paper has identified a number of challenging research issues. Although progress has been made, we are still a long way from being able to build a functioning geolibrary, for reasons that are both technical and institutional. The GIS research community has much to offer in both areas.

References

GOODCHILD, M.F. and K.M. DONKIN (1967) A computerized approach to increased map library utility, *The Cartographer*, **4**.

HAWKINS, B.L. (1996) The unsustainability of the traditional library and the threat to higher education. Paper presented at *the Stanford Forum for Higher Education Futures*, The Aspen Institute, Oct. 18.

MORRIS, B.A., BARTLETT, D.J., DOWERS, S., GITTINGS, B.M., HEALEY, R.G., IRVINE, J.J.A. and WAUGH T.C. (1986) Cartographic information retrieval and the creation of graphic indices from database records. Paper presented at the *British Library Research and Development Department Symposium*, Cranfield Institute of Technology, 27 July, Department of Geography, University of Edinburgh.

SMITH T.R., ANDRESEN, D., CARVER, L., DOLIN, R., GOODCHILD, M.F. and others (1996) A digital library for geographically referenced materials, *Computer*, **29** (5), 54+ (corrected author list appears in **29** (7), 14).

CHAPTER SIX

The KINDS project: providing effective tools for spatial data accessibility and usability over WWW

ADRIAN MOSS, JIM R. PETCH, A. HEPTINSTALL, K. COLE, C.S. LI, K. KITMITTO, M.N. ISLAM, Y.J. YIP AND A. BASDEN

6.1 INTRODUCTION

The mere existence of rich data resources does not guarantee that they are accessed and used effectively. Effective use is dependent upon awareness, accessibility and usability of data resources. Also, to be able to use a data set users must be appropriately trained. Many spatial data sets are large and complex, requiring expert skills and knowledge. Raper and Green (1992) attribute an acute shortage of spatial data handling skills to a lack of awareness of the potential usefulness of spatial data, the need to learn concepts associated with the new technology and to develop specific technical skills. The skills shortage is compounded because existing skills are not readily transferable and retraining may take 18 months to three years (Davies and Medyckyj-Scott, 1994).

These factors can restrict data-use to a relatively limited number of technically highly-experienced users. The problem is exacerbated by potential users who, lacking the necessary skills, may even be unable to evaluate properly the usefulness of a spatial data set. This can create a self-perpetuating cycle which restricts the number of spatial data users to those already experienced. Potential new users may be reluctant to invest time in learning how to use spatial data handling tools before ascertaining whether the available data sets suit their needs.

The Knowledge-based Interface to National Data Sets (KINDS) Project is a response to these problems. Its aim is to reduce the overhead of using spatial data sets by tackling three key problems: low *awareness* of the content, coverage and structure of the data sets; poor *accessibility* of data, which is held in different formats and manipulated using different tools; and low levels of data *usability* which requires users to be equipped with the knowledge they need to use data effectively (Petch *et al.*, 1997). KINDS has developed a series of interfaces on the world wide web (WWW) to address these problems. This paper comprises three sections. In the first, the user-orientated methodology employed