FUTURE DIRECTIONS FOR GEOGRAPHIC INFORMATION SCIENCE

Michael F. Goodchild

National Center for Geographic Information and Analysis, and
Department of Geography
University of California
Santa Barbara, CA 93106-4060, USA
Phone: +1 805 893 8049; FAX: +1 805 893 7095
Email: good@ncgia.ucsb.edu

Abstract

Geographic information science (GISci) can be defined as the set of basic research issues raised by the handling of geographic information. Although geographic information systems are often seen merely as tools, there is ample historical precedent for the role of tools in stimulating science, and provoking new ways of thinking about problems. GISci is a distinct specialty within a more broadly defined information science, with a multidisciplinary base that ranges from surveying to cognitive science. The paper addresses two topics among the many potentially fruitful areas for research within GISci at this time. It traces the development of debate in the U.S. over the National Spatial Data Infrastructure, and its implications for the research and education communities. Spatial data provides a distinct set of problems for the development of digital libraries, and the paper reviews the Alexandria Project and some of its more fundamental objectives. The paper concludes with comments on the current status of GISci within the sciences generally.

INTRODUCTION

Some years ago I published a paper entitled "Geographical Information Science" (Goodchild, 1992). At the core of the paper was a play on the acronym GIS, which we normally assume to refer to geographic information systems. In part, I wanted to respond to comments that seemed to marginalize GIS as "a mere[1] tool". Jordan, for example, had referred to GIS as "non-intellectual expertise" in a presidential column in the Newsletter of the Association of American Geographers (Jordan, 1988). The history of science provides ample evidence that there is nothing "mere" about many scientific tools. The microscope and telescope both produced an outpouring of new science and new ways of thinking about the world. By comparison, however, the effects of the digital computer have been immeasurably greater, not only in science but in the structure of society as a whole, and we are only just beginning to see the impacts of the revolution in information technology that is still occurring. I wanted to suggest that when a

[1] The four-letter words "just", "mere", and "only" must be among the most powerful in the English language in their ability to imply marginality; there seem to be no such simple terms to imply centrality.

researcher in the GIS community used the acronym, they were probably referring to something more fundamental than a class of software in its current, limited level of development. Since the paper appeared the term seems to have attracted some attention. Last December a group of about 30 U.S. universities took the first steps to forming a University Consortium for Geographic Information Science, and Texas A&M University recently announced that they would offer a degree under that name.

The 1992 paper identified geographic information science (GISci) as the set of fundamental scientific issues arising from, stimulated by, or surrounding the use of digital computers to handle, process, analyze, store, or access geographic information. Within that domain, it is easy to find issues ranging from the modeling of uncertainty in geographic information (a problem in spatial statistics), to the design of user interfaces (cognitive science), to the design of algorithms, data models, and data structures (computer science), to the representation of real geographic phenomena (geography?). The last example is of such importance that one might go so far as to argue that the greatest challenge in GISci is to find ways of building useful representations of the infinitely complex world around us in the almost absurdly limited, discrete environment of a digital computer.

Information is rapidly becoming the most important commodity of human economic activity. The days when trade was dominated by agricultural products are long gone, and industrial products are rapidly losing ground in international and interpersonal transactions. It is now possible for information about an industry to be worth as much or more than the industry itself (Niemann and Niemann, 1993). The transition to an information-based economy creates enormous problems for us, as well as opportunities. We have seen an explosion of new information technologies in the past few years, and it is already difficult to remember the days before Email, when it could take days to prepare and send a letter. To most of us, the Internet is less than five years old, and it is hard to believe that the growth of the World Wide Web has all occurred in the past two. Compared to these time scales, our institutional frameworks seem impossibly dated, and governments, educational institutions, and legal systems seem doomed to an ever more hopeless game of catch-up. The idea that one can plan and order the future of human society cannot possibly survive under the strength of this onslaught from the information technologies.

Unfortunately, education seems to be among the most conservative of industries. Our students receive very little systematic education in the information sciences, despite their evident and increasing importance in the real world. Surely information - its theory, collection, access, storage, manipulation, value, legal context - should be a focus of education from the earliest years. Yet the notion persists that information science is not real science - it has no experiments, and it hardly fits with the popular Hollywood image of the wild-eyed scientist in the white coat beside the bubbling beaker of steaming, colored fluid.

One of the more difficult issues we face concerns the market value of geographic information. Many information industries have faced the valuation problem by requiring purchasers to buy all of the available data, or one of a limited selection of well-defined subsets.

Analog geographic information has often been bought and sold by map sheet. But digital geographic information is in principle seamless. Should it be bought and sold by byte, or by square meter of coverage? In practice, most exchange seems to occur in tiles, such as map sheets or counties or image scenes, perpetuating an analog practice that has no basis in the digital world. It seems that the seamless 1m resolution panchromatic imagery about to be marketed commercially by various consortia will use an interesting compromise - it will be sold in fixed quantities that roughly correspond to the near-square area covered by a conventional aerial photograph, but the user will be able to define the exact position of the area on the Earth's surface.

Despite issues like these, the new information technologies present unparalleled opportunity, particularly for the kinds of information that have always been difficult to create, find, access, and analyze. Within this context, what directions will geographic information science need to take if geographic data is to take full advantage of the new technologies? The next sections of this paper address two areas that seem to contain the kinds of issues that need to be tackled by a geographic information science. Although there are clearly many more, these two are of particular interest at this time, as they correspond to major areas of current interest in the U.S. The first of these sections discusses the National Spatial Data Infrastructure (NSDI), and the second efforts to realize the concept of a digital library for spatial data.

THE NATIONAL SPATIAL DATA INFRASTRUCTURE

In the past few years, NSDI has emerged as the umbrella for a series of initiatives and discussions aimed at a coordinated national spatial data policy. NSDI falls within the broader umbrella of the National Information Infrastructure (NII), a national effort to exploit modern technologies for electronic communication. The term "information superhighway" suggests an analogy to the national efforts beginning in the 1950s to stimulate the U.S. economy through the construction of a national freeway network. NII's best known manifestation to date is the Internet, but it includes many other private and local networks.

The geographic data community is certainly not unique in attempting to identify its own piece of NII; biologists, for example, have proposed a National Biological Information Infrastructure. The focus of NSDI on geographic data makes sense if one recognizes the significance of geography to the functions of government, particularly in the development and implementation of policy for such areas as the environment or health, and particularly if significance is measured by volume. It reflects a recent assessment of the total annual government expenditure on geographic data, which found $4.5 billion of activity in 14 departments. Perhaps it also reflects a genuine new interest in geography, visual communication, and spatial thinking, and in making locational information as easy to deal with as any other kind.

The definition of the NSDI contained in a report of the Mapping Science Committee of the National Research Council is one of the most often quoted:

"The means to assemble geographic information that describes the arrangement and attributes of features and phenomena on the Earth. The infrastructure includes the materials, technology, and people necessary to acquire, process, store, and distribute such information to meet a wide variety of needs" (National Research Council, 1993).

Although NII is most often thought of as the technical infrastructure of the Internet, NSDI places at least as much emphasis on the process of coordination, in the form of standards, educational and training programs, and partnerships between different levels of government and with the private sector as it does on technical matters.

At the federal level, NSDI activities are coordinated by the Federal Geographic Data Committee (FGDC), a group representing the policy-forming level of management within 14 different federal agencies, all concerned to varying degrees with the production and use of geographic information. FGDC was mandated by Circular A-16 of the Office of Management and Budget (OMB), a department of the executive branch of the federal government; it is staffed by the U.S. Geological Survey, and currently chaired by the Secretary of the Interior. Also relevant to NSDI at the federal level is OMB Circular A-130, which directs federal agencies to share data with others at the cost of reproduction. At the state level, much NSDI activity is coordinated through the National States Geographic Information Council (NSGIC).

Much of the impetus behind NSDI comes from the support it has received from the Clinton Administration. The report of the Gore Commission (National Performance Review, 1993) committed the Administration to building the NSDI in collaboration with state and local governments and the private sector. An Executive Order signed by the President in April 1994 laid out the short term objectives of NSDI activities in the federal government, including the establishment of a National Geospatial Data Clearinghouse, identification of the data sets that will form the framework of NSDI, and adoption of a standard for metadata that would help to make geographic data easier to find and share by providing a formalized means of description, allowing potential users to determine a data set's fitness for their own needs.

The framework is seen as a coordinated collection of digital data sets that will allow users to build their own data and applications on a quality-assured foundation. It is likely to include the geodetic control network, which provides the highest quality of submeter georeferencing with respect to an established set of fixed monuments. At a somewhat coarser level, it might include a digital cadaster, but that is unlikely in the near term given the cost of producing such a coverage for the nation, and the nature of the U.S. land ownership system. It is likely to include topographic data, by adopting the largest scale mapping available in any area; this currently varies from scales around 1:2,000 in some urban areas to 1:24,000 in rural areas of the continental U.S., and 1:250,000 in parts of Alaska, providing the potential for georeferencing to approximately 1m, 12m, and 125m respectively. Digital road networks are also likely to be included, since they provide the most convenient source of georeferencing in urban areas with accuracies of perhaps 12m, depending on the source.

The newest addition to the set of potential framework data sets are digital orthophoto quads (DOQs). In their commonest form, these offer raster panchromatic images with a pixel size of 1m, and a positional accuracy of approximately 6m, planimetrically corrected for topography and camera angle. DOQs lack any identification of features, but are nevertheless useful for a variety of purposes, including mapping of soils and vegetation, and updating of topographic maps. Perhaps their greatest use is as a framework for digitizing specialized features, such as road or utility networks. Their far lower cost allows for more frequent update than topographic maps, and they offer significant advantages as a contribution to the NSDI framework.

NSDI is seen as a truly national effort, despite the fact that much of the leadership currently originates in the federal agencies. The framework will be built through partnerships, to nationally agreed standards. Because the scales of much data will vary depending on the source, particularly for the topographic data sets, NSDI will make heavy use of automated methods of generalization.

Although many aspects of building the NSDI may seem straightforward, in practice there are likely to be substantial impediments that will require coordinated research. First, the geographic information community has struggled for decades with the problem of cataloging maps and images, and the move to digital has made the problem even more difficult. Although an FGDC metadata standard is now in place, it is not likely to be the final answer, and there are still serious gaps in our theoretical understanding of the nature of geographic data on which a comprehensive standard would have to be based. Second, NSDI calls for the heavy use of automated generalization techniques that do not yet exist; we are still far from being able to automate the complex system of rules and constraints developed by cartographers (Buttenfield and McMaster, 1991).

NSDI envisions a world of digital geographic information in which standards will make it easy for data to be transferred from one system to another, and the promulgation of the Spatial Data Transfer Standard, now known as Federal Information Processing Standard 173 (Morrison, 1992), has moved us a long way toward this goal. But the world of format standards will still require considerable effort on the part of the user to assemble and reformat data from various sources, perhaps with projection change and resampling as well. Recently, much interest has been expressed in the concept of interoperability, which would allow the user to access data without becoming involved in issues of format. In this object-oriented world, data would be encapsulated with the processes needed to transform it. The Open Geodata Interoperability Specification (OGIS) is attempting to lay the ground rules for GIS interoperability, with the cooperation of many of the major GIS vendors.

Besides research, the goals of NSDI will require significant investment in the educational system, to expose the next generation of users of geographic information to its principles and components. NSDI data should be available in the school classroom, to provide school students with the framework they need as they explore the world around them, and build knowledge about it. Much progress has been made toward introducing GIS to the schools, but the sheer size of the problem is daunting.

Finally, while NSDI represents a very significant move toward a coordinated approach to geographic information in the U.S., there is a pressing need to extend its principles to the international and global arenas. Estes and Mooneyhan (1994) have recently commented on the surprisingly poor state of the georeferencing framework for the globe, and its possible impacts on research into global systems and global environmental change. Efforts are under way to improve this situation, but it will take a massive and well coordinated effort if it is to succeed.

DIGITAL SPATIAL DATA LIBRARIES

Libraries provide access to information, but the traditional library limited itself largely to printed information in the form of books. The advent of the digital catalog made the task of finding books much easier, and online catalogs available over the Internet have made it possible to search for books without actually visiting the library--of course, one still needs to visit to read or withdraw the book.

Digital technology has now advanced to the point where it is possible to consider placing the books themselves online. Electronic journals already exist, and circulate freely over the Internet without ever existing in print. An average university library contains perhaps 10^{13} bytes of information in printed text form, and that amount of information could be stored in a single file store using currently available technology. Of course the cost of data input would be enormous, but so also would be the benefits. The library would be fully accessible to anyone connected to the network, and it would be possible to take advantage of entirely new capabilities for searching text, or following hypertext-like links between topics.

Despite their successes with books, traditional libraries have never been as comfortable with geographic information, in the form of maps and images. They are difficult to catalog and cumbersome to handle, and often relegated to "special libraries" where most users find the presence of a human assistant essential in interpreting their needs and finding suitable materials. Few successful digital catalogs exist for maps, so users must almost always visit the library in person even to establish that a certain map is stored there.

In a digital world, many of these problems of dealing with geographic information might be expected to disappear, allowing geographic data to move into the library mainstream, and to be fully merged with other types of information, including text, video, and sound. Data volume is a potential problem, as a large map library arguably contains an order of magnitude more bytes of information than a large conventional library. But geographic information should be as easy to find and access, and perhaps even easier to interpret.

For this goal to be achieved, however, certain significant impediments need to be removed, and they lie within several domains of GISci. First, the typical map library user accesses information through the use of set of keys that is potentially much more complex than the familiar author/title/subject keys of books. Of these, location is almost always the most important, followed by some combination of scale, date, theme, and possibly several other characteristics. However, the location key can be expressed either through a coordinate system, by finding the area of interest on a map, or through a gazetteer by finding the appropriate placename. The scale key can be expressed in at least four distinct ways, as the metric scale of the document used to create the digital data set, or as a pixel size, or a spatial resolution, or as a measure of positional accuracy. Some of these may even be undefined, but may serve as convenient surrogates for others, as when "scale" is used as a surrogate for the contents of a topographic database. Themes are often combined in complex ways on maps, and it is difficult to construct useful thesauri of geographic data themes.

In early 1994, the U.S. National Science Foundation (NSF), in cooperation with the Advanced Research Projects Agency (ARPA), and the National Aeronautics and Space Administration (NASA), solicited proposals for six research projects in digital libraries. One of the successful proposals was the Alexandria Project, a collaboration between the National Center for Geographic Information and Analysis, several departments at the University of California, Santa Barbara, and several corporations and libraries. It is the only project of the six focused explicitly on geographic data and other geographically referenced materials such as books about places, or photographs. Information about Alexandria can be found on the World Wide Web at http://alexandria.sdc.ucsb.edu/. NSF funding is for four years, during which time the project is expected to produce a testbed implementation of its research. In the short term, we have produced a "rapid prototype" system based on ESRI's ArcView II, Sybase, and the interface scripting language Tcl/Tk, and populated it with sample data sets drawn largely from UCSB collections and concentrating on the Santa Barbara region. The prototype allows the user to zoom in on a base map, define a search window, query the metadata base for all materials covering areas that intersect with the search window, and retrieve the contents of any of those materials.

In the next few months, we plan to create a Web presence for the Alexandria prototype, to allow Internet access. The current prototype lacks certain obvious requirements, including a gazetteer and thesaurus, and these will be added in the next development period. We also plan to tackle the problems of scale and date specification, by building appropriate user interface tools. Over the next four years we plan to grow from an initial focus on spatial data to a full digital library designed for all information types.

In the longer term, the concept of a digital spatial data library presents several fundamental challenges to GISci. While it is straightforward to deal with queries and materials with crisp rectangular footprints aligned with the cardinal directions, we also need to deal with queries about areas that have fuzzy edges, and with materials that have fuzzy geographic extents. We currently lack a comprehensive set of models for fuzzy geographic extents, or methods for working with them at the user interface level.

Second, we must deal in a comprehensive way with the issue of content-based search. Consider a map library user who requires information on the Mississippi River. Unless that specific feature appears in the title of a library item, which seems unlikely, he or she will have to rely on knowledge or experience as a guide to the maps likely to contain it. In a digital environment, the retrieval will have to be based not on metadata, but on the contents of the library's items. Moreover, the material retrieved may be more useful if it is in the form of an image rather than a feature-based data set. Because the image representation will not identify the feature explicitly, it will be necessary to find it through some kind of linkage between a feature-based dataset and a raster image. This will require an approach that is very different to the organization of the traditional map library or imagery collection, or even of today's spatial database or GIS.

Libraries have been organized around the principle that information is stored in discrete units, either as books or as journal articles. Similarly, map libraries manage and store discrete map sheet units, or discrete air photos. This basic granularity of information allows catalogs to be structured as records with homogeneous formats. Hierarchical structures are comparatively rare, although they occur in map libraries when some metadata fields are stored for individual maps, and some as general characteristics of entire map series.

In the digital library, these ideas of information granularity are likely to disappear, or be replaced by more general concepts. The Microsoft *Encarta* encyclopedia, for example, is not a book or a map or a journal in the conventional sense, but a mass of different types of information with complex relationship structures. It uses novel concepts in its user interface that have no analogs in the traditional library. *Encarta* and its competitors offer some useful pointers to how the user will interact with the digital library of the future. The first generation of digital spatial data libraries will probably be close analogs of the traditional library. But future generations are likely to adopt very different approaches - seamlessness, new metaphors of data exploration such as the "magic lens", and new methods of indicating the presence of information about areas on the Earth's surface.

CONCLUDING COMMENTS

The previous two sections have discussed only a very small selection of the possibilities currently offered by GISci. They happen to be two that I am familiar with through my own research. Both include fundamental issues, and hard research problems, and both are strongly motivated by meaningful applications, characteristics that seem broadly true of this field. GISci seems at an early stage of growth, judging by the wealth of new problems it presents.

When NCGIA began in 1988 one of the missions assigned to it by NSF was to promote the use of geographic information analysis and GIS within the sciences generally. At that time it seemed that GIS had made inroads into disciplines like geography, but other fields had yet to discover it or use it in their research. It offered the potential to remove many of the impediments to looking at data from a spatial perspective, an approach that is clearly useful for any discipline dealing with phenomena distributed over space.

Since then there has been enormous progress, and GIS is now widely accepted in many disciplines. There are regular conferences on GIS and environmental modeling, GIS and hydrology, GIS and anthropology, and GIS sessions in many disciplinary conferences. As the field has matured, the needs have changed. Today, GIS finds itself in somewhat the same position as statistics - there is a similar need for a discipline that coordinates research into geographic information and analysis, helps in its applications, and provides the materials on which education programs can be based. We face a universal reaction from the application disciplines that GIS is too difficult to use; one can imagine a similar reaction from users faced with understanding statistical packages without a basic course in statistics. Our campuses are not likely to create departments of geographic information science, so we must find new ways of providing coordination, research, and help, from an interdisciplinary base. This is one of the arguments behind the foundation of the University Consortium for Geographic Information Science in the U.S., and it is an argument that is likely to be heard frequently in the next few years.

Modern science is interdisciplinary. Earth system science, for example, requires understanding of the coupling between atmosphere and oceans, and close coordination between atmospheric scientists and oceanographers. Interdisciplinary science requires the sharing of data and tools, components of the research enterprise that have traditionally been given much less importance than science itself. The information sciences like GISci can flourish along with this new way of doing science, as the ability to handle, store, describe, exchange, and organize information becomes increasingly the key to solving some of our most pressing problems.

ACKNOWLEDGMENT

The National Center for Geographic Information and Analysis and the Alexandria Digital Library are supported by the National Science Foundation, Grants SBR 88-10917 and IRI 94-11330 respectively.

REFERENCES

Buttenfield, B.P., and R.B. McMaster, 1991. *Map Generalization: Making Rules for Knowledge Representation*. Harlow, UK: Longman Scientific and Technical.

Estes, J.E., and D.W. Mooneyhan, 1994. Of maps and myths. *Photogrammetric Engineering and Remote Sensing* 60(5): 517-524.

Goodchild, M.F., 1992. Geographical information science. *International Journal of Geographical Information Systems* 6(1): 31-45.

Jordan, T., 1988. *AAG Newsletter* 23(5): 1.

Morrison, J.L., 1992. Implementing the Spatial Data Transfer Standard: Introduction. *Cartography and Geographic Information Systems* 19(5): 277.

National Performance Review, 1993. *Creating a Government that Works Better and Costs Less*. Washington, DC: U.S. Government Printing Office.

National Research Council, 1993. *Toward a Coordinated Spatial Data Infrastructure for the Nation*. Washington, DC: National Academy Press.

Niemann, B.J., and S.S. Niemann, 1993. GIS as a megatrend: lunch with John Naisbitt. *Geo Info Systems* 3(4): 64-67.