Michael F. Goodchild
National Center for Geographic Information and Analysis, and
Department of Geography
University of California
Santa Barbara, CA 93106-4060

TECHNICAL ADVANCES IN SPATIAL DATA SHARING

Abstract: NCGIA's Initiative 9 was conceived largely before today's methods of electronic communication became widely available and popular. New Internet-based tools and new institutional frameworks have vastly altered the environment of digital spatial data sharing over the past eight years by providing electronic implementations of traditional data sharing concepts. The digital library brings all of these together in an electronic analog of the traditional functions of the library. The paper ends with a discussion of the Alexandria Digital Library project and its focus on spatially indexed information.

INTRODUCTION

The title and broad subject matter of NCGIA's Initiative 9 were set in late 1987, when the consortium members met to finalize the list of research topics to be included in the proposal submitted to the National Science Foundation in early 1988. Although several years elapsed, the research agenda developed at the initiative's Specialist Meeting in early 1992 was still broadly consistent with the original concept. Since then, however, we have seen enormous progress in the technologies needed to make sharing possible, and increased attention to sharing of data not only between institutions but within the scientific community, and in society generally. The ability to share data and information easily and cheaply is now seen as one of the major goals of the information age. Much spatial data is now available using technologies that were almost inconceivable in 1987, resulting in dramatic changes in the context of spatial data sharing. Moreover, we can now begin to see some of the special problems that spatial data raises, and that make the effective sharing of spatial data more than usually difficult to achieve.

In 1987 the technical cutting edge of spatial data sharing was a 2400 ft reel of 9 track tape in the mail. Bitnet was becoming popular for electronic mail, particularly amongst users of certain mainframe operating systems, and although ARPANet was widely used in certain research communities the term "Internet" was virtually unknown. Today, the technical cutting edge uses client/server

software operating over the Internet, offering the prospect of unlimited access to a worldwide data resource. This paper provides a review and summary of current technology for data sharing, and the new institutional arrangements that have emerged with it. It then moves to a discussion of the outstanding issues, and new ones that will have to be solved in the near future to help us move closer to the eventual goal of universal, cost-effective spatial data sharing.

CURRENT STATUS

Data sharing requires several stages, and the existence of technical solutions to each of them. A need must be established by a user, who must then find a way to express that need and conduct a search to identify the existence and location of suitable data resources. A custodian of data must publish information on its existence and availability in a fashion that is appropriately informative. The published data must be compared to the requirement. If a fit is found, the required data must be retrieved from storage and transmitted by the custodian. Issues such as payment, preservation of intellectual property rights, and security must be resolved. Finally, the user must read the data, perhaps also translating or reformatting it to make it readable. Although all of this may be conducted in a digital world, it includes analogies to several traditional processes. Technical progress in digital data sharing over the past eight years has relied heavily on the ability of electronic systems and networks to provide these digital analogs.

The first of these was surely the development of an electronic analog of mail. Bitnet was the first system to provide a widely available capability for sending mail-like messages from one computer user to another. It could be used to search for data, by sending messages to likely custodians asking if data was available, and to transmit data by including it as part of the message. The development of list servers and other methods of broadcasting messages to multiple users allowed for a crude form of publishing by supporting the dissemination of information on data availability, and the electronic analog of junk mail became a problem.

The Unix operating system also encouraged direct computer to computer communication, by providing appropriate tools to manage operations over wide area networks. By using telnet, ftp, or rlogin, it became possible for the transmission of data from custodian to user to be initiated and controlled by the user. Although this may seem to be a minor advance from a technical perspective, it is very significant sociologically, since it allows or empowers the user to function independently of the custodian. The user need not ask the custodian for permission, and the custodian no longer derives power from the possession of data.

Internet emerged as a loose confederation of regional networks in the late 1980s, and quickly began to outgrow its competitors. In part Internet's success is the result of its reliance on minimal but widely acceptable protocols, such as

TCP/IP, and the lack of any rigid institutional hierarchy; in part it is due to Internet's origins in the research community, and the goals of its funding agencies. In the past few years a series of new applications and tools have appeared on the Internet that implement analogs to several other traditional processes, and thus move the technical basis of data sharing forward at a rapid pace (a range of general reviews of the Internet and its applications are available; see, for example, Braun, 1994; Dern, 1994).

First, the Internet now supports various forms of search. Tools such as WAIS first encourage custodians of information to publish its existence according to standard protocols. They then allow the user to initiate a search across all published information, using a simple key word strategy. Various metrics have been developed to indicate to the user the likelihood that any given collection of information will suit the requirements of the search. Suitable information can then be retrieved by the user. Such systems rely on the willingness of custodians to publish the availability of data, but otherwise leave control entirely in the hands of the user.

A second development has revolutionized the Internet since its release in early 1993. The World Wide Web (WWW) and its user interface tools such as Mosaic provide an analog of the traditional process of browse. But whereas traditional browse is often linear, as when a library user scans shelves, or a reader scans an encyclopedia, WWW exploits a more powerful concept of direct logical linkages between information sources, either within one site or across the Internet. Its second major advantage is its use of graphics and color to enrich information. The casual user can access and retrieve a vast array of published information including photographs, maps, and images, at no cost, with minimal training, and with the lowliest home computer. Finally, it empowers custodians by giving them sole control over what they will make available; and it empowers users by giving them sole control over what they will look for and retrieve. The Mosaic client software is sufficiently simple that it could be made available simultaneously and free for all major personal computer architectures, and would not need to be frequently updated; and the server software is simple enough to allow almost anyone to make their information available.

Because of their simplicity and almost universal adoption over the past two years, Mosaic and WWW have been used to make a wide range of spatial data available for sharing. For example, the Montana State Library provides access to its spatial data holdings through its Maps of Montana project (http://nris.msl.mt.gov/gis/mtmaps.html)[1]. The Distributed Spatial Data Library maintained by ERIN (Earth Resource Information Network) in Australia is a particularly extensive and impressive source

[1] This Universal Resource Locator (URL) provides the Internet address needed to access the relevant server.

(http://kaos.erin.gov.au/dsdl/dsdl.html). Through WWW links, it is possible for a single source to provide "virtual" access to a much larger network of other sources.

Many of these implementations use the "point and click" capability to provide links to data. The GLO/EPA Wetlands Resource Database Project (http://www.texas.gov/DIR/gis-image.html), for example, presents users with a map of Texas and allows them to specify areas of interest by pointing to red dots on the map. More sophisticated queries, such as a user-defined polygon of interest, are obvious extensions for finding spatial data but not possible with current protocols and client software.

In summary, recent advances in network communication and applications have revolutionized the context for spatial data sharing. The next section looks briefly at changes in the institutional context.

NATIONAL SPATIAL DATA INFRASTRUCTURE AND SHARING

The concept of a National Spatial Data Infrastructure (NSDI) has emerged almost simultaneously with the growth of the Internet, and has become an umbrella for U.S. spatial data activities within the National Information Infrastructure, or "information superhighway" (NRC,1993). The Executive Order signed by the President in April 1994 assigned responsibility for NSDI at the Federal level to the Federal Geographic Data Committee (FGDC), and committed the Administration to building NSDI.

Data sharing is a primary motivation behind NSDI. The National Research Council report (NRC, 1993) argued that substantial duplication of effort existed within the U.S., as different agencies created digital representations of the same data, and for a variety of reasons failed to share them. It was difficult to find out what data existed, because of a lack of widely accessible catalogs and means of description; standards were inadequate and often conflicting; and no mechanisms existed for identifying those data sets most suitable for sharing.

Within the past year, FGDC has made great strides in developing the first stages of NSDI. A National Geospatial Data Clearinghouse has been developed to provide information on the availability of data - a "virtual" catalog. A metadata standard has been developed and adopted as a framework for describing the characteristics of data to potential users. And progress has been made on identifying "framework" data sets to form a shareable basis of data for NSDI.

The primary purpose of identifying framework data sets, and making them available in digital form over the Internet as a common resource, is clearly to promote the concept of sharing, and avoid duplication of effort. The data sets most likely to be shared are those that fill a generic purpose, and allow users to add their own information. The strongest candidates for framework data sets are

those that support georeferencing, by allowing users to position features and phenomena relative to other, well-defined features of known position. Obvious candidates as framework data sets include digital versions of the geodetic control network, digital topographic maps, digital street centerline data, and digital orthophoto quads and elevation models. In the present political and economic climate it is clear that these data sets will have to be produced through partnerships between the private and public sectors, including all levels of government (NRC, 1994), and the private sector will be particularly important in cases where economic activity provides sufficient incentive, as is clearly the case with street centerline data.

Effective data sharing relies on the ability of the custodians to describe data in ways that are meaningful to potential users. The term "metadata" has largely replaced the more traditional "catalog" in this context, and also conveys some of the sense of the term "documentation". It is often defined as "data about data", suggesting that one test of whether an item of information is metadata is whether it is true of all or a substantial proportion of the records in the data. Obvious items of metadata include the vertices of the bounding polygon of the data set; dates of creation and ground validity; scale or spatial resolution; and quality. Since metadata is a property of a data set, it should travel with the data, and be updated when the data is transformed, such as through a change of projection.

Producers and custodians of spatial data tend to have strongly held concepts of data granularity, or the appropriate contents of a data set. In many cases these derive from traditional approaches to map production, or are determined by data modeling choices. For example, it is easy to think of the contents of a map sheet as a unit of spatial data, with associated metadata. TIGER files have often been managed by county, and Landsat data is often managed by scene. But while these concepts of granularity may make good sense from the point of view of the producer, or the librarian, they are often in conflict with contemporary spatial database concepts of seamlessness and layering. From a theoretical perspective, it may make more sense to approach the design of metadata from a perspective of seamlessness. This issue will be central in the design of distribution systems for the new remote sensing sources, particularly the high resolution imagery that will be appearing in the next few years. Should users be required to deal with predefined "tiles" with fixed footprints on the Earth's surface, or will they be able to request data for arbitrarily defined areas, and if so how will the value of data be assessed - by unit area?

THE LIBRARY METAPHOR

Traditionally, one of the most important institutions in the distribution of information has been the library. It has acted as a warehouse, storing a substantial quantity of published information; as a place to browse, or search for information to meet a defined need; as a method of exploiting economies of scale

by allowing users to share access to material rather than purchasing it themselves; and as an institutional framework for cataloging. From the perspective of communication, it functions as an intermediary between producer or publisher and user, by accepting or purchasing information from the former and lending it to the latter. In the context of our earlier discussion, the publication is the only significant step in the traditional chain from custodian to user that is not merged into the functions of the library.

Spatial information presents problems for traditional libraries. Maps and images are cumbersome, requiring special methods of storage such as map cabinets and humidity control. Even more problematic are the difficulties associated with cataloging. The methods of cataloging books - by subject, author, and title - are inherently discrete, since it is possible in principle to make a list of finite length of all possible authors, titles, and subjects. Moreover, all three indices are readily sorted by alphabetical order or through hierarchical schemes such as the Library of Congress catalog. In contrast, the primary index for accessing maps and images is location, and virtually all queries to map libraries specify this as the most important characteristic. But location is multidimensional and continuous, and it is impossible in principle to make a finite list of all possible locations. Fixed tiles, such as the Landsat scene boundaries or USGS quadrangles, provide only partial solutions for some items, and there is no single, obvious approach to sorting them.

For these and other reasons maps and images are not normally part of the traditional library. They may be relegated to a special map room, and the lack of standard methods of cataloging and difficulty of access are reflected in the high ratios of staff to users found in these facilities.

Libraries began widespread adoption of digital technology in the 1970s, with the use of computers to handle catalogs and improve capabilities for search. The idea of digitizing the materials themselves - the books, journals, and other items - is comparatively recent, and reflects vast improvements in the price and performance of large-volume digital storage, using such media as optical disk. Moreover, there is every reason to expect continued improvement. The digital library is conceived as a system in which all information - catalog, books, journals - is in digital form, allowing vastly improved capabilities for search and retrieval, as well as access and retrieval in fully distributed fashion over the Internet, allowing any user to access any information from a workstation, without ever visiting a physical library.

In 1994, a consortium of three Federal funding agencies - NSF, ARPA, and NASA - announced a competition for six four-year grants to support the research that would address the technical issues involved in building digital libraries. Unlike many countries, the U.S. has never had a National Library, and the idea that this might be achieved through digital technology with universal access is a significant part of the vision of this and related programs.

In the digital world, many of the traditional problems associated with libraries of spatial information should disappear or become much less significant. Map information is no more cumbersome to store in a digital database than any other kind of information, although it is comparatively voluminous if volume is measured on the basis of traditional map and image granules. Digital approaches should also make it possible to deal effectively with the problem of continuous indexing. One of the successful projects funded under the digital library program is Alexandria, an effort centered at the University of California, Santa Barbara, within the Map and Imagery Laboratory of the Library, and including researchers from the Departments of Geography, Computer Science, and Electrical and Computer Engineering. All three sites of NCGIA are participating in various aspects of Alexandria. The project's main purpose is to focus on the special problems of maps, images, and spatially indexed materials, and to ensure that they are fully addressed in the digital libraries of the future. In effect, we hope to bring spatial data back into the information mainstream.

The Alexandria design includes analogs of all of the traditional library functions - ingest, catalog, store, and retrieve. It uses an approach to metadata that is compliant with both the FGDC standard for spatial metadata and the library MARC standard. Our plan is to provide for distributed access over the Internet, to distributed information servers linked by common protocols, data formats, and methods of description.

Spatial data - maps and images - provide the initial motivation for Alexandria. However, a host of other types of information can be spatially indexed, and the ability to access such information through the spatial dimension may be very valuable. For example, one should be able to find a travel guide to Paris this way, or a geologic guide to the State of Nevada. Historical photographs of early Los Angeles may be much more useful if spatially indexed, and accessible through a map or similar indexing scheme. A historian may be interested in all materials in the library related to a particular city or region, and a student of literature may wish to link novels through their common geographical setting.

At the highest level, Alexandria must provide a conceptual framework for all types of information, since its eventual goal is to integrate spatial and spatially referenced data into this universal framework.

The cooperative agreement with NSF required us to produce a prototype of Alexandria within the first six months. Our Rapid Prototype (Frew et al., 1995) was constructed using commercial, off-the-shelf software (ESRI's ArcView, Sybase, and Tcl/Tk). It allows the user to specify an area of interest on an appropriately scaled base map, search a file of metadata records for maps and images that overlap the area and satisfy any other specified criteria, display a "thumbnail" of any selected "hits", and retrieve the data set using one of several recognized native raster and vector data formats.

KEY ISSUES FOR DIGITAL SPATIAL DATA LIBRARIES

The objectives of Alexandria raise a number of issues of fundamental significance for the sharing of spatial data, particularly if such sharing is to occur through the implementation of the library metaphor. The issues of continuous versus discrete indexing, and granularity have already been addressed. The selection of materials by location may occur through the use of a map, by allowing the user to point to an area of interest, or outline a query polygon. It can also occur through the use of a gazetteer, or list of standard geographical names, and a map library user is actually far more likely to identify an area of interest by name than by coordinates. It is clearly necessary to support both methods of access. Most gazetteers, such as the Geographic Names Information System (GNIS) of the USGS, provide only a point reference for a given feature, but some concept of extent will be needed if the system is to find and display the feature appropriately scaled on a base map. We are developing a tool set for specifying extents, and plan to implement this along with the gazetteer.

One of the issues we will have to deal with is the "fuzzy" extent. While the Rapid Prototype handles rectangular query areas aligned with the cardinal directions, some users may be looking for data on areas that have no well-defined, crisp footprints, such as "downtown Denver". We will have to support fuzzy query areas, and also fuzzy extents for named features in the gazetteer. This will require the development of a set of standard templates for describing fuzzy areas, such as polygons with blurred boundaries, or blurred circles. A catalog of easily defined templates with economical mathematical descriptions is an important first step.

Another set of issues relates to what is often termed "content-based search" (CBS). Consider a user asking for Landsat scenes showing the Mississippi River. Since most approaches to metadata focus on information that is true of all or much of the dataset, it is unlikely that specific features covered by the data set will appear in the metadata, no matter how prominent, except in cases where a prominent feature is the basis of a data set title. One way to handle this issue is to allow for metadata fields that can be defined by specific application domains, and perhaps filled by automated pattern recognition. For example, a geologist might be interested in knowing which images include bedrock outcrops, or a meteorologist might be interested in having the presence of hurricanes identified in metadata.

Topographic maps are designed to identify features which are of generic interest to many applications; are well-defined in the field; and are therefore useful in identifying position. This suggests that we can achieve something approximating content-based search if we build tight linkages between images and feature-based data sets. The Mississippi River query could be satisfied by retrieving topographic coverage at an appropriate scale; finding the feature and its extent; and finally retrieving the appropriate image. Of course if the feature

exists in the gazetteer the process can be shortened considerably. Thus one key element in our work plan for Alexandria is to reexamine the relationship between image-based and feature-based data sets, and to link them much more effectively than has been the tradition in the past in spatial data handling.

The concept of scale also raises some fundamental issues. Although most cartographers use it routinely to characterize digital data sets, metric scale loses any rigorous meaning once data is digitized, and becomes a surrogate for other characteristics such as spatial resolution, positional accuracy, size of smallest object, or database contents. Remote sensing specialists tend to prefer pixel size as a surrogate for spatial resolution. Spatial resolution is a confusing concept to many users of spatial data, and tends to be referenced indirectly or by implication, if at all (Frank et al., 1995). Alexandria will have to provide intelligent assistance in helping users specify their spatial resolution requirements.

WWW and Mosaic provide an obvious and readily available means for establishing Alexandria facilities on the Internet. However, the current version uses a simple "point and click" approach that severely limits the user's ability to interact with visual presentations, and makes it impossible, for example, to specify a query polygon. The first Internet implementation will probably use the current version of Mosaic, and more advanced versions are likely to become available in the lifetime of the Alexandria development.

CONCLUDING COMMENTS

One of the more intriguing questions underlying Alexandria concerns the role of space as a means of access to information. Today, our ability to use a map to find information is limited to the map sheet indexes of map libraries; indexes in some atlases; and innovative tools on the Internet such as the Virtual Tourist. Projects such as Alexandria make it possible to find maps and images using a map, but also raise the possibility that maps might become much more widely used and effective methods for finding information of all kinds. But whether a map will ever be useful to someone looking for a copy of Agatha Christie's "Death on the Nile" remains to be seen - an author catalog seems more likely.

NCGIA's Initiative 9 overlapped a period of intense advances in the technology of data sharing, and was able to benefit from the massive stimulation of interest in data sharing that occurred as a result. We now have much better tools for overcoming the problems of duplication and redundancy that led to the definition of the initiative eight years ago; we have viable institutions in place to foster data sharing; and we have much greater awareness of the benefits that can result. More importantly, perhaps, we also have an awareness of the specific and unique problems raised by the sharing of spatial data, as distinct from other data types. Spatial data presents unique problems of granularity, indexing and cataloging, and assessment of fitness for use, and we have made progress in

solving these in the past few years. We still have a long way to go, however, in developing methods for finding data to satisfy specific needs. Most specific needs for spatial data are only approximately satisfied. Technically, one might see degree of satisfaction as assessed by a metric computed by comparing the user's metadata request with the custodian's metadata response. Such metrics will be increasingly important as the amount of metadata to be searched explodes, and will have to be designed specifically for spatial data searches. At this point, such metrics remain an unsolved issue.

REFERENCES

Braun, E., 1994. The Internet Directory, New York: Fawcett Columbine.

Dern, D.P., 1994. The Internet Guide for New Users, New York: McGraw-Hill.

Frank, S., M.F. Goodchild, H.J. Onsrud, and J.M. Pinto, 1995. "Framework data sets for the NSDI", unpublished research report, National Center for Geographic Information and Analysis, University of California, Santa Barbara.

Frew, J., M. Aurand, B.P. Buttenfield, L. Carver, P. Chang, R. Ellis, C. Fischer, M. Gardner, M.F. Goodchild, G. Hajic, M. Larsgaard, K. Park, M. Probert, T.R. Smith, and Qi Zheng, 1995, "The Alexandria Rapid Prototype: building a digital library for spatial information", unpublished paper, Alexandria Digital Library, University of California, Santa Barbara.

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

National Research Council, 1994. Promoting the National Spatial Data Infrastructure Through Partnerships, Washington, DC: National Academy Press.

ACKNOWLEDGMENT

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