FINDING GEOGRAPHIC INFORMATION: COLLECTION-LEVEL METADATA

Michael F. Goodchild[1]
Jianyu Zhou[2]

Abstract

Certain types of information are associated with specific locations on the Earth's surface, and can be retrieved using location as the primary search key. The combination of a collection of such information and a location-based search mechanism is termed a geolibrary. A search for specific information over multiple collections is more efficient if one knows where to look. Information about the contents of a collection is termed collection-level metadata (CLM). Several conceptual designs for the process of search are reviewed, and the U.S. National Geospatial Data Clearinghouse is discussed. The Alexandria Digital Library provides an example of a large digital collection. Its geographic coverage is analyzed to determine whether its CLM can be modeled effectively. The difficulties of describing CLM are discussed from the perspective of institutional change, as library collections attempt to make themselves more accessible through the Internet.

Introduction

Although the Earth's surface is seamless, geographic data nevertheless tend to occur in finite aggregations commonly termed data sets, files, or databases. In this paper we focus on data sets that relate to well-defined areas or *footprints* on the Earth's surface. For example, a single remotely sensed image or *scene* from a sensor such as Landsat's Thematic Mapper has a footprint that is roughly rectangular, and on the order of 100km on a side. The concept of footprints is not limited to maps, images, and their digital counterparts—data sets that represent variation over the Earth's surface and that we commonly term *geospatial*—but also includes books, reports, photographs, and many

[1] 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 3146. Email: good@ncgia.ucsb.edu.
[2] Department of Geography, University of California, Santa Barbara, CA 93106-4060, USA. Email: zhou@geog.ucsb.edu.

other forms of information that describe some aspect of a geographic area and thus have well-defined footprints (Goodchild, 1998). The concept can also be extended to the fuzzy case (Burrough and Frank, 1996; Montello *et al.*, 1998), where it is not always precisely clear whether a given point lies inside or outside a given footprint, especially if the point lies close to the footprint's boundary. Footprints are useful characteristics of datasets because they enable search mechanisms based on geographic location (libraries with such search mechanisms have been termed *geolibraries*; Mapping Science Committee, 1999).

In this paper we refer to such data sets as *geographic information-bearing objects* (GIBOs), and assume for the purposes of discussion that they are unlikely to be either subdivided or aggregated. This basic granularity of geographic information is of course largely arbitrary, the result of production mechanisms and many other factors, but it is unlikely to change fundamentally in the near future. A large number of such data sets exist, in many cases grouped into *collections* for purposes of access. Search is then a two-stage process, requiring first an identification of an appropriate collection, and then a search within that collection for the needed data set. This two-stage model is applicable both to the traditional library and to the virtual libraries of the Internet. In the traditional library case, a library's collection of information-bearing objects is assembled gradually by the library's acquisition and collection-building processes, and each object in the collection is catalogued, shelved, and made accessible through the library's search mechanisms. In the Internet case, the *site* is roughly equivalent to the library, and identified by a single point-of-entry (often a *page* or WWW universal resource locator, or URL). The various information-bearing objects offered by the site are accessed from this page by the site's search mechanism, which may take various forms but will have been designed to make the user's task of locating information as straightforward as possible.

In this paper we focus on the first stage of the process. In the Internet case much progress has been made on the second stage, that is, the process of finding information within a site, notably in the National Geospatial Data Clearinghouse project (*http://www.fgdc.gov*, and see the Geospatial One-Stop at the same site), in research prototypes such as the Alexandria Digital Library (*http://www.alexandria.ucsb.edu*), Microsoft's Terraserver (*http://terraserver.microsoft.com*), the MIT orthophoto server (*http://ortho.mit.edu*), and many others. But the first stage—or the question of *how one* *knows where to look*—remains comparatively unexplored. In the traditional library it was resolved by a simple heuristic: *all research libraries contain all important information*, and thus reduced to the task of finding any research library. Since most users visit traditional libraries in person, it is appropriate to choose that research library that minimizes the associated travel, by visiting the closest research library or the one accessible at the least transport cost. Goodchild (1997, 2001) has pursued this concept of the research library as a central service for a geographically dispersed population of users, by embedding it in the framework of central facilities location theory, or central place theory (Berry *et al.*, 1988). Research libraries tend to be ranked based on the degree to which the heuristic works, in other words, the likelihood that if one visits the library one will indeed find all important information; and also to rank information based on its general importance. But Goodchild argues that both of these principles fail in the case of geographic information, and more generally in the case of what he terms *information of geographically determined interest* (IGDI), which tends to be of more importance in areas within or near the information's footprint, and therefore more likely to be found in certain libraries than in others.

Consider the question "Do you have information on the life of the person who invented the theory of natural selection?" Ten years ago one might have gone to the nearest research library and put the question to a librarian. A conversation might have ensued, in which the librarian suggested modifications to the request to fit the information known to be in the library, or made the request easier to accommodate (*e.g.*, "Are you thinking of Charles Darwin? If you are, you can look up biographies in the card catalog by the name of the subject person. Or I could find the book for you once I've finished this.") Today, the search for information is likely to be conducted on the Internet, with its vastly improved ability to provide access to information to anyone who is electronically connected to its networks. But whereas the first stage of the search was often trivial in the library case, it is far from trivial in the dispersed, uncatalogued Internet.

Now consider the IGDI case, and the question "Do you have information on clear-cuts in Maine?" If the nearest research library was in Los Angeles or London, it is very unlikely that anything useful would result—instead the user might go away empty-handed, or with the promise that the request would be pursued through some form of inter-library loan, perhaps by sending it electronically to the library at the University of Maine. In the Internet case there are no librarians, and it is the user's responsibility to pursue the request alone. Moreover the Internet contains on the order of 10^7 sites, some of them located in or near the State of Maine. The needed information might exist anywhere, though geographic proximity is a reasonable basis for guesswork. But it might also exist at sites associated with governments or agencies at any level in the geographic hierarchy: a state government site, a county site, a non-governmental organization's site, or the personal site of a local researcher.

The purpose of the paper is to identify an emerging problem associated with distributed resources of geospatial data: the lack of an efficient basis for guiding a search over potential sources. The paper explores the magnitude of the problem, and various potential solutions. The paper is structured as follows. In the next section we define the problem formally. In the third section we discuss current solutions to the problem, and their advantages and disadvantages, within that formal framework. In the fourth section we define collection-level metadata (CLM), and in the fifth section describe an experiment to generate it. The final section presents further discussion and our conclusions.

Problem formulation

Consider a dispersed population of servers, each connected by the technology of the Internet and able to provide certain services. Each server is managed by a *custodian*, who defines and maintains the set of information objects mounted on the server, and builds the necessary search mechanisms. Let the number of such servers be N (currently on the order of 10^7), and let them be indexed by i. Some of these servers contain the needed information, the number of such potential *hits* lying between 0 and N. We assume for the purposes of this discussion that the search ends in success if at least one hit is found, and in failure once it is established that the number of potential hits is 0. Thus the second stage of the search process, of locating the needed information within a hit, and the subsequent stages of retrieval, manipulation, or analysis, are not relevant here: for our purposes it is sufficient to establish that the site is a hit.

In the worst case all N sites must be tested, and the amount of work that must be performed depends directly on N. Define the metric $\log N$ as a suitable measure of this work, or of the inaccessibility of the information. To reduce the amount of work we must somehow reduce the effective value of *N*. In the extreme case where all information is accessible through one site, the effective value of *N* is 1 and log *N* is reduced to zero. In the traditional library the card catalog provides such a one-stop solution, organized to maximize the likelihood that a user will be able to satisfy a request through this one source of abstracted information.

Suppose now that some prior estimate exists of the probability that a site *i* contains the information; denote this probability by p_i. We could interpret *p* in a frequentist mode by suggesting that in a search over *n* sites with the same value of *p* the expected number of hits is *pn*; or the interpretation of *p* could be strictly subjective. The search is trivial if p_i=1 for at least one *i*, and will not be conducted if p_i=0 for all *i*. We cannot assume that the number of hits is limited to 1, or that the sum of p_i over *i* equals 1. But the belief that some sites are more likely to contain hits than others clearly changes our expectations about the amount of work that will have to be undertaken. If no p_i=1, a modified information statistic:

$$H' = -\sum_i p_i \log\left[\frac{p_i}{\sum_i p_i}\right] \tag{1}$$

provides a suitable measure of inaccessibility; if at least one p_i=1 we define *H'*=0. *H'* has a maximum value of log*N* when all *p* are equal and sum to 1, and is suitably increased when all *p* are equal and sum to more than 1. *H'* tends to zero if one server is much more likely to contain a hit than other servers, and generally behaves like the information statistic in measuring the distribution of *p*.

H' is defined over a set of *N* servers, and measures the inaccessibility of information located on one or more of them, given no knowledge about the information's location except prior probabilities. By contrast, traditional methods for measuring accessibility in the geographic world, such as one might use to measure the accessibility of a library, focus on the deterrent effects of distance, rather than on the size of the set of possible sites (see Janelle and Hodge, 2000, for a review of the concept of accessibility in real and virtual worlds). For example, a common measure of accessibility of a location to a set of sites indexed by *i*, and where A_i measures the attractiveness or level of service of the *i*th site, is the Potential Index:

$$V = \sum_i {A_i}/{d_i} \tag{2}$$

where d is the distance to the ith site (Olsson, 1965). But in the world of the Internet the deterrent effects of distance are reduced almost to zero, and inaccessibility is defined not by the costs of travel but by uncertainty about a site's contents.

In this paper we focus on the definition of p. GIBOs are IGDIs, and most likely to be of interest in areas within or near their footprints. Although in principle it would be possible to serve the demand for a GIBO from any server, independent of the server's location, since the costs of overcoming distance are close to zero on the Internet, in practice we expect a GIBO to be found on a server within or near its footprint for several reasons (Goodchild, 1997, 2001). First, government agencies at all levels are more likely to produce GIBOs for areas within their own jurisdictions, and a similar if weaker effect may also apply to the private sector. Second, geographic proximity between footprint and custodian tends to reduce the costs of maintenance and update. Third, there remain significant distance-related impediments to interaction on the Internet, despite the electronic speed of transmission. The bandwidth available between two places tends to decrease with the distance between them, particularly at global scales, as does the observed latency of interaction (Dodge and Kitchin, 2001; Murnion and Healey, 1998). Thus we expect p_i, the probability that a given GIBO will be found on server i, to depend to some extent on the geographic distance from the GIBO's footprint to the server.

In summary, we propose to measure the likelihood that a server i contains a hit by p_i, and to measure the magnitude of the search task with a modified information statistic H'. The value of H' can be reduced if methods can be found to estimate p, since H' is reduced if one or more p are higher than the others, and reduced to zero if at least one p_i=1. In this paper we deal specifically with the IGDI case, and the use of geographic location as a basis for estimation of the p_i.

Current strategies

If H' measures the magnitude of the search task, then it is clearly desirable to adopt strategies that are likely to reduce its value. A number of such strategies are in widespread use, and others have been prototyped; we review them in this section.

Clearinghouses

In the clearinghouse solution (Figure 1), analogous to the union catalog often maintained by a group of libraries, one server acts as a one-stop shop by providing a within-site search mechanism, and either offers the objects itself, or provides URL links to servers containing the actual information objects. Within the geospatial domain a well-known example is the U.S. National Geospatial Data Clearinghouse (NGDC). GIBOs on over 250 servers world-wide are accessible through this site (*http://www.fgdc.gov*), which is currently mirrored at 6 clearinghouse sites around the U.S. (the user is directed to select the "nearest" of these sites). The clearinghouse pages allow the user to formulate a query, and to select specific servers from among the list of over 250. For this scheme, p_{NGDC} is the probability that a given information object is accessible through the clearinghouse. In the ideal world, where all geospatial data is accessible through this one site, H' reduces to 0 for all geospatial data. In practice, access through NGDC depends on the willingess of the custodian of a given server to register its existence with NGDC, to adopt NGDC's metadata rules, to conform to the Z39.50-based protocol on which the clearinghouse operates, and to complete the necessary procedures. To estimate p_{NGDC} for a given GIBO, a user must estimate the likelihood that the custodian of the GIBO participates in NGDC and has made the GIBO accessible through this mechanism. In practice this is based on a number of highly *ad hoc* and learned heuristics, such as the knowledge that public sector agencies, particularly agencies that are members of the U.S. Federal Geographic Data Committee (FGDC), are more likely to make use of the clearinghouse, along with many states. But from a quick scan of the NGDC pages it is clear that participation is very difficult to predict; it would be hard to guess, for example, that significant resources of South African geospatial data are accessible through this mechanism.

Eventually, NGDC may succeed in attracting participation by a large proportion of custodians of geospatial data. Until that happens, H' will remain substantially greater than zero. Moreover, it seems likely that other clearinghouses will emerge, perhaps at the state level sponsored by state governments, or perhaps sponsored by libraries or mission-oriented agencies, to compete for the attention of custodians. The developer of a clearinghouse is likely to encourage registration, because arguments for continued funding can be based on use, and use is likely to depend on the volume of data registered with the clearinghouse. Private sector data may be less likely to appear in clearinghouses, which may assume or even require that all data accessible through them be in the public domain (but compare ESRI's Geography Network, *http://www.geographynetwork.com*, which provides access to both public-domain and commercial data resources). Custodians of GIBOs that are not geospatial, or for which geospatial content is comparatively unimportant, may find the protocols of the geospatially oriented NGDC difficult to comply with, and may instead register with clearinghouses that do not support geographic search. For various reasons, then, and despite its evident success, the clearinghouse model seems unlikely to provide a comprehensive solution to the search problem.

Search engines

The search engines of the WWW also attempt to reduce *H'* to zero by providing a one-stop shop for information using automatic or semi-automatic processes. Search engines send out large numbers of automatic agents that detect the existence of pages, identify significant keywords using a series of sophisticated strategies, and provide an index to pages using those keywords. In effect, search engines build and support the digital equivalent of card catalogs, substituting URLs for the shelf numbers of library books. But their success is limited by several factors. First, the rigorous structure of author and title that exists for all published books has no equivalent for material mounted on WWW servers. Instead, search engines must examine the contents of information and attempt to identify the most characteristic keywords, a process that is often successful but essentially unreliable. Second, the WWW has no equivalent of the library's subject catalog. Subject catalogs make use of a shared and formally recognized *controlled vocabulary* or *authority* of subjects that is used by all library cataloguers and accessible to all users, coupled with a *thesaurus* or list of synonyms that can be used to determine relevance when keywords do not occur in the vocabulary. Third, it is possible that the important subjects do not occur in text in the object, and thus cannot be detected by search engines. This occurs frequently in the case of GIBOs, since it is common for footprints to cover areas that correspond to no obvious and widely recognized text place name. Since any part of a GIBO's footprint may be the focus on search, it would be necessary for an object to be identified by all of the placenames within the footprint, or perhaps overlapping with it. Conversely, a GIBO that covered only part of the search footprint might be useful and worth accessing.

In summary, Internet search engines as currently designed are not successful at searching by geographic location, because search footprints are rarely expressible in the words of a placename. Instead, footprints are often more rigorously defined by latitude and longitude, which are measures on a continuum rather than words from a controlled vocabulary. Search engines have no way of comparing numerical values, and cataloging on that basis, especially when these are multi-dimensional.

It would be relatively easy to fix this problem if the Internet community could be persuaded to adopt a simple protocol for tagging objects by geographic footprints, and if search engines could be designed to recognize and process these. The proposed Dublin Core (*http://www.oclc.org*) and other similar efforts might be successful in providing the necessary protocol, but there remains the question of the willingness of custodians to comply with any new standard.

There is also the possibility of designing agents with sufficient intelligence to recognize GIBOs, and to infer their footprints by recognizing standard geospatial data formats, coordinate systems, placenames, or other indicators. MapFusion, software developed and marketed by Global Geomatics Inc. (*http://www.globalgeo.com*), searches a defined network domain to identify files matching any of several hundred recognized geospatial data formats; opens the metadata heading such files; builds a catalog; and allows users to open the files. Like WWW pages, files must be given suitable permissions (*e.g.*, located in public_html directories) if they are to be found by this process.

Hybrid solutions

Dolin *et al.* (1997) describe Pharos, a multi-level hierarchical approach to the problem. It proposes a large number of duplicated high-level (local) catalog servers, and a smaller number of mid-level servers, to direct queries to appropriate sources (Figure 2). A user would first query the local server, and would be directed to one or more of the mid-level servers based on matching the characteristics of the query to known specialties. For example, one mid-level server might be known to specialize in material about South

America. The mid-level servers would possess more detailed metadata, and would direct the user to specific sources, where the most detailed metadata are found. But H' is still finite since neither the local nor the mid-level servers have complete knowledge of source contents.

NGDC fits this model to some extent, if one equates the mirrored local servers of Pharos with the 6 mirrored sites of NGDC. But there are no mid-level servers in the NGDC scheme, and the user must specify the sources to be searched. It is difficult to do this effectively, except by guesswork, based on the names of the sources. Thus "Massachusetts Coastal Orthophotos" is a relatively precise name for a source that serves a very specific type of data for limited coverage; "State of Kansas Data Access and Support Center" would be a reasonable choice for a query involving Kansas; but in a case like "Cornell University Geospatial Information Repository" it is difficult to make an effective choice without more information.

Collection-level metadata

Although other potential solutions can be identified, for the time being it is clear that a user searching for information must rely on straightforward search guided by prior estimates of the probability that a given site contains appropriate information. The typical experienced GIS user possesses an enormous amount of information of this type, and shares it or guards it as the case may be. We may remember a recent email announcement that all DEMs for a certain state are now available at a given URL, or that a certain university project recently collected a large amount of geospatial information for a given area, or that a catalog of geospatial data was recently produced by a certain state government. Or we may make use of email lists to circulate requests for knowledge about information. All of these informal methods redefine the set of p_is and reduce H' to varying degrees.

We define *collection-level metadata* (CLM) as information about the contents of a collection. The term *collection* has a range of meanings (Hill *et al.*, 1999), but here is defined simply as a group of objects accessible through a common catalog, such that the existence of a given item of information in the collection can be determined unambiguously. Both the objects and the catalog might be stored on a single server; or the catalog might exist on one server, and point to objects on many servers. CLM summarizes the *object-level metadata* (OLM) that is the focus of such standards as the FGDC's Content Standard for Digital Geospatial Metadata (*http://www.fgdc.gov*; see also ISO 19115), and thus guides the search among servers, as OLM guides the search among information objects on a single server (or on multiple servers connected by protocols, as in the case of a clearinghouse).

Because it summarizes OLM, CLM might contain any of the fields defined for OLM. But instead of specific values, in most cases CLM must provide distributions. For example, a server of digital topographic data in vector format might include some objects at 1:10,000 scale, some at 1:25,000, and some at 1:50,000. At the object level, each metadata record would contain a single entry for scale. But the CLM must include information about the collection as a whole. The more accurately the CLM summarizes the OLMs, the more likely p_i will be either 1 or 0, obviating the need to search the OLM records in order to resolve whether the server contains a hit. Hill *et al.* (1999) describe the relationship between CLM and OLM for collections in the Alexandria Digital Library, and identify two types of CLM, which they term *inherent* and *contextual*. Inherent CLM can be derived through automated analysis of a catalog, and includes distributions of objects by temporal and spatial coverage, and by type. Contextual CLM is supplied by the collection provider or maintainer, and cannot otherwise be derived automatically from content; it includes representations of the collection's scope and purpose, and its protocols and schemata.

Clearly CLM works best when it consists of simple rules that accurately summarize content. *All* and *none* are effective bases because they uniquely define underlying OLM. For example, a server that contains *all* DOQs is much more useful than one that contains *some*, unless some additional rule defines *some*, such as *all DOQs for Massachusetts*. The most difficult collections to summarize are those that include *some* objects in a given class, and where the other qualities of those objects also vary. For example, the CLM for a collection of assorted topographic maps, of varying scales and dates, acquired over many years from miscellaneous research projects, would be almost impossible to express effectively at the collection level.

Various projects have implemented CLM, and many of these occurred before the advent of digital libraries and the WWW. Hill *et al.* (1999) review many of these *finding aids*. The STARTS protocol is a recent attempt to improve the performance of network search engines using various forms of CLM to describe the contents of servers and to achieve interoperability between OLM catalogs (Gravano *et al.*, 1997; Lagoze, 1998).

CLM and the Alexandria Digital Library

The Alexandria Digital Library (ADL) is one of six projects initiated under the NSF/DARPA/NASA (the U.S. National Science Foundation, Defense Advanced Research Projects Agency, and National Aeronautics and Space Administration, respectively) joint Digital Library Initiative of 1994. Its objective is to provide the services of a map and imagery library over the Internet, using the Map and Imagery Laboratory (MIL) of the University of California, Santa Barbara, as the working example. Several prototypes have been built since the project was initiated, and the most recent is accessible via *http://www.alexandria.ucsb.edu*.

Like any prototype geolibrary, ADL assumes that the primary basis of the user's search is geographic location. Two methods of defining location are supported. A user can select placenames from a *gazetteer* (for a review of gazetteers and their role in geolibraries see Hill, 1998). In the current implementation the ADL gazetteer includes order 10^7 names corresponding to locations around the world. ADL then converts the placename to a query footprint expressed in global coordinates using the services associated with the gazetteer. Alternatively a user can interact directly with a map of the world, panning and zooming to define a query footprint directly in global coordinates. After defining the query footprint, the user can refine the search by specifying a combination of additional properties of GIBOs, including subject, level of geographic detail, date, producer, etc. ADL implements a subset of the FGDC OLM standard, harmonized with the U.S. MARC standard of the library community, and currently has order 10^6 GIBOs represented in its collection and OLM.

Since location is the primary key for search in a geolibrary, it follows that footprints are the most important components of both OLM and CLM, and the primary basis for determination of p_i. In this paper we discuss the analysis and modeling of the locational component of ADL CLM, in an effort to provide a source of estimates of p_i, or the probability that information related to a certain footprint is available in ADL. Such information will be of great value to a user in deciding whether to search the ADL OLM catalog for specific needs. More general issues associated with ADL CLM are discussed by Hill *et al.* (1999). In principle the ADL architecture can accommodate many collections, but in practice it is the collection of GIBOs that is of interest (ADL's architecture also manages its gazetteer as a "collection").

The ADL collection, which amounts to order 10^6 objects, has arisen as a result of a series of priorities and compromises. There are two major components to the cost of entering or *ingesting* an object into the ADL collection: the cost of conversion of the object to digital form, and the cost of creation of the OLM; both require varying levels of human intervention. In practice, objects already in digital form have been favored, as have objects that already possess OLM in digital form. In addition, the ADL collection has been guided by the contents of the MIL, which is located on the UC Santa Barbara campus and emphasizes coverage of the region local to Santa Barbara, as well as complete coverage of certain national map and data series, and objects of historical value such as early aerial photographs and satellite images. The contents of the collection have been driven in part by a series of repository agreements with map- and image-producing agencies.

Each object in the ADL collection has a footprint, defined as a pair of latitudes and longitudes, in other words a rectangle aligned with the axes of a cylindrical projection in equatorial aspect. It is an easy matter to create a bivariate histogram of these footprints, and an example is shown in Figure 3. We created a regular grid using equal divisions of latitude and longitude, and counted the number of ADL objects having a non-zero intersection with each cell. Note that objects that intersect *n* cells are counted *n* times, and an object that covers the entire surface of the Earth is counted in every cell. Note that this histogram summarizes only the locational component of CLM; a full CLM would have to present a multivariate histogram that contains all other practically important search keys, such as scale or date, since it is not possible to assume in general that the probability of the collection possessing a data set for a defined area is independent of the probability of it possessing a data set for a defined scale or date.

Modeling CLM

In order for a user to assess *p* it is necessary for CLM, such as that represented in Figure 3, to be readily accessible. In the clearinghouse model, for example, a representation of the geographic coverage of the collection accessible through the clearinghouse would have to be available to the user; earlier we discussed the current arrangement, which assumes that the user of the clearinghouse somehow already possesses a mental representation. In the Pharos model, a digital representation of Figure 3 would be available at each mid-level server. A more abstracted version, such as a simple indication of which major region of the Earth's surface was best represented in a collection, would also be available at each local server. The amount of information that would have to be distributed in such a scheme to the mid-level and local servers depends on the degree to which Figure 3 can be represented by a simple model. For example, if each collection can be characterized accurately by identifying the major region that is the focus of the collection, then a simple code would suffice as a representation. Clearly CLM that can be presented in the form of a simple rule, for example by stating that the collection contains all available information on Massachusetts, would be preferable to a complex picture such as that presented in Figure 3, or to a complex index map. Such simple rules would provide values of p_i close to 0 or 1, whereas Figure 3 shows that p_is for the ADL collection are likely to lie very far from 0 or 1. We have analyzed the geographic distribution of the ADL collection to demonstrate the difficulty of finding such a simple rule.

Earlier, we argued that such collections represent a legacy of the heuristic "every research library contains every object". But because traditional libraries require their users to visit them in person, and most users visit the library that is closest to them, the density of users declines rapidly with distance from the library. And because geospatial data are IGDI, the geographic coverage of a collection will to some degree reflect the geographic locations of its users. Thus we expect the coverage of a collection also to decline with distance, a pattern that is clearly evident in Figure 3.

The literature of economic and behavioral geography describes many models of human behavior over distance. Wilson (1970) and others showed from maximum-likelihood (maximum entropy) considerations that the probability of a user visiting a central facility should decline as a negative exponential function of distance, and models of the following form are used widely in practical applications such as retail choice modeling (Fotheringham and O'Kelly, 1989; Haynes and Fotheringham, 1984):

$$P_{ij} = aE_i A_j \Big/ e^{-bd_{ij}} \tag{3}$$

where P_{ij} is the probability that a user in area i will visit facility j from among the available facilities, a and b are constants, E_i is a measure of the propensity of area i to generate usage, A_j is a measure of the propensity of facility j to attract usage, and d_{ij} is the distance from area i to facility j.

With a single facility we drop the j subscripts. In our application the areas i are the cells defined by latitude and longitude, and since they vary in area we include that factor in the model, to allow for larger areas near the equator and smaller areas near the poles. Geospatial data is likely to be more abundant for areas of land than for sea, so we include a binary variable defined by the majority of the surface area in the cell. We measure distance using a spherical model of the Earth. Finally, since some areas of the Earth's surface are better-mapped and likely of more widespread interest than others, we include population density as a variable in the model:

$$Y_i = b_0 A_i^{b_1} D_i^{b_2} e^{b_3 d_i + b_4 L_i} \tag{4}$$

where Y_i is the number of objects covering the ith cell, A_i is the area of the cell, D_i is the population density of the cell, L_i is 1 if the majority of the cell's area is land, else 0, and b_0 through b_4 are constants to be determined.

Calibration of (4) using non-linear methods could be problematic, especially if use of small cells results in a very large sample. Fortunately it is easy to linearize the model by taking logs, and this also has the advantage of helping to stabilize the variances. The transformed model has the form of a conventional linear regression model:

$$\log Y_i = \log b_0 + b_1 \log A_i + b_2 \log D_i + b_3 d_i + b_4 L_i \tag{5}$$

Since Y represents a count of events, we anticipate that it will have an error distribution that is Poisson with a variance proportional to its value; again, taking logs will help to stabilize the effects of unequal variances.

Data preparation

From spherical trigonometry we determined the surface area of the cell bounded by longitudes λ_1 and λ_2 and latitudes φ_1 and φ_2 as $R^2(\lambda_1\text{-}\lambda_2)(\sin\varphi_1\text{-}\sin\varphi_2)$, where R is the Earth's radius, and computed A from this expression. We determined d by computing the great circle distance over a spherical Earth from the center of each cell to the location of the library. We used a radius R=6378.137m. We obtained L by rasterizing a vector database of global boundaries compiled by GRID-Geneva, a center of the U.N. Environment Program. All data sets were registered to the same grid defined by latitude and longitude. We used a 1 degree cell in the following analysis, with a potential sample size of 64800.

We obtained D from the data set compiled by NASA Goddard Institute for Space Studies, and identified as the "Global Distribution of 1984 Population Density at 1 degree by 1 degree resolution." The data set was compiled by Inez Fung, Elaine Matthews, and their associates. After rasterization, approximately 2/3 of the 64800 cells were water with zero population density, and many land cells were recorded as uninhabited. In addition, some parts of the world were assigned missing codes in the NASA data because data had never been compiled for them by any government or non-government organization. We treated these cells also as having zero population density. A total of 9886 cells had positive population densities.

Calibration

We first fitted the linearized model (5). We dealt with cells where D_i=0 (log0 is negative infinity) by the arbitrary device of assigning them values of 1 (log1=0). The linear model gave a multiple R^2 of 0.560, and the coefficients are shown in Table 1.

The cases of D_i=0 presented us with a dilemma. On the one hand model linearization is desirable both because it allows the model to be calibrated by ordinary least squares, and because it helps to stabilize variances. On the other hand the device of adding 1 is clearly arbitrary, and has an effect on regression results. For comparison, therefore, we also calibrated (4) using the nonlinear regression procedure available in Splus. It gave the coefficients shown in Table 1. Note, however, that direct comparison of R^2 is difficult because the dependent variables are different [log$Y$ for (5) and Y for (4)].

Coeff	Vble	Model 3 value (linear model)	Model 3 std. error	Model 2 value (nonlinear model)	Model 2 std. error
b_0		3.13	0.014	294.0	55
b_1	*A*	0.179	0.0015	-0.0954	0.021
b_2	*D*	0.0425	0.0015	0.126	0.0030
b_3	*d*	-0.0000500	0.00000028	-0.000142	0.0000012
b_4	*L*	0.204	0.0033	0.244	0.031

Table 1: Linear and nonlinear regression results for the ADL collection

The distance variable *d* has a negative effect, as expected: the likelihood of the ADL collection holding a given object declines with distance between the library and the object's footprint. *L* has a positive effect, coverage of land area being approximately 23% more likely in the linear case and 28% more likely in the nonlinear case, all other things being equal (to be equal, for example, ocean would have to be compared with land of zero population density). Population density has the expected positive effect. The effect of area is more problematic, however. All other things being equal, one might expect a coefficient of 1, since coverage should scale linearly with the area of the cell. But area covaries with latitude, and therefore with numerous other effects.

We noted earlier that ADL regards its gazetteer as a form of collection. More broadly, placenames occur in many library catalog records, in book titles or in subject classifications, and it is possible to analyze any library's digital catalog for the geographic distribution of such placenames. The result might be thought of as the geographic CLM of the collection. Similarly, the geographic coverage of ADL's gazetteer reflects varying interest in different parts of the world; there is dense coverage of North American placenames, and much less dense coverage of other areas. To demonstrate this, Figure 4 shows a 1 degree summary of ADL's gazetteer coverage, for comparison with Figure 3. To compute it, we assigned every entry in the gazetteer to its containing 1 degree cell using the entry's representative point, and ignored any information on the geographic extent of the feature where this was available.

Table 2 shows the comparable analysis of the gazetteer data, using the same models (4) and (5). The linear model gave a multiple R^2 of 0.683.

Coeff	Vble	Model 3 value (linear model)	Model 3 std. error	Model 2 value (nonlinear model)	Model 2 std. error
b_0		-2.67	0.054	10.9	2.5
b_1	A	0.399	0.0060	0.152	0.025
b_2	D	0.392	0.0018	0.346	0.0031
b_3	d	-0.0000493	0.0000011	-0.000131	0.0000012
b_4	L	1.10	0.0132	0.259	0.026

Table 2: Linear and nonlinear regression results for the ADL gazetteer

Distance shows a similar negative effect, as expected. The land/water distinction is more pronounced than in Table 1, reflecting the comparative lack of named places in oceanic areas.

Discussion

The question that drove the previous section was whether it is possible to model the coverage of a collection of geographically referenced information, such as one finds in the Alexandria Digital Library. As a form of CLM, such a model would be a valuable tool in helping users determine where to look for information. The results of the experiment are decidedly mixed: coverage declines with distance from the library, and is greater in areas of high population density and on land, but the effects are weak, and the guidance that would be given to the user by this form of CLM would be highly unreliable.

ADL's collection is somewhat arbitrary, as noted earlier, having resulted from a large number of compromises over a long period of time. In other cases, such as the MIT orthophoto server identified earlier, the descriptions “coastal Massachusetts” and “orthophotos” provide a precise description of content, and would be much more successful as CLM. But this is a relatively small collection assembled for a specific purpose. Many of the servers now participating in NGDC have collections more like ADL’s—somewhat miscellaneous in coverage, theme, source, and other characteristics, and offered to a largely undefined population of potential users in the hope that some of them may find some of the data useful. It seems to us that many efforts to provide data over the Internet will resemble ADL’s, and that as the number increases the problems associated with imprecise CLM will grow worse, unless comprehensive solutions can be found.

One of the objectives behind the development of ADL was the desire to make the resources of a world-class map and imagery collection accessible to a larger number of users. From the perspective of the central facilities model, the Internet was seen as an ideal mechanism for overcoming the impeding effects of physical distance, which had previously confined use of the collection to those living within a short distance of the library. The library’s collection reflected the needs of its users, and also followed the heuristic identified earlier: all major research libraries contain all important information. But as we have argued, the essential question of where to look had a simple answer: look in the nearest research library.

ADL would have succeeded in this objective if it had been able to digitize its entire collection and to provide the necessary access tools, and thus to become a one-stop-shop for georeferenced data. But in practice it was necessary to select certain items for digitizing, based on assorted criteria that include levels of use and intellectual value, because of limited funds. Thus the ADL collection is a poorer approximation to the heuristic “all important information” than its parent. Moreover, in a world of multiple servers and universal access it is unnecessary for all servers to serve all information, provided some basis can be found for knowing where to look. Thus attempting to replicate a traditional library collection in a digital collection is wrong for two reasons: first, it is unlikely that funds will support digitization of the entire collection; and to do so under conditions of universal access would result in vast duplication of effort—the equivalent of moving all research library collections to a single building. Instead, what is needed is a fundamental shift of institutional arrangements, from one that has evolved to replicate the contents of research libraries, to one designed to produce unique and narrowly defined collections with precise CLM.

In the absence of precise CLM, it seems to us that the only possible solution is a new generation of search engines, designed to identify and catalog the footprints of data sets and to support search. This would be aided if data sets could be given recognizable tags in some appropriate format, such as already occurs in the NGDC. Intelligent agents could be designed to recognize certain common geospatial data formats, placenames, and coordinate systems, to populate appropriate metadata records, and to make them searchable.

References

Berry, B.J.L., and J.B. Parr, with B.J. Epstein, A. Ghosh, and R.H.T. Smith (1988) *Market Centers and Retail Location: Theory and Applications*. Englewood Cliffs, NJ: Prentice Hall.

Burrough, P.A., and A.U. Frank, editors (1996) *Geographic Objects with Indeterminate Boundaries*. London: Taylor and Francis.

Dodge, M., and R. Kitchin (2001) *Mapping Cyberspace*. New York: Routledge.

Dolin, R., D. Agrawal, A. El Abbadi, and L. Dillon (1997) Pharos: a scalable distributed architecture for locating heterogeneous information sources. *Proceedings, Sixth International Conference on Information and Knowledge Management, CIKM '97*. http://pharos.alexandria.ucsb.edu/publications

Fotheringham, A.S., and M.E. O'Kelly (1989) *Spatial Interaction Models: Formulations and Applications*. Dordrecht: Kluwer.

Goodchild, M.F. (1997) Towards a geography of geographic information in a digital world. *Computers, Environment and Urban Systems* 21(6): 377–391.

Goodchild, M.F. (1998) The geolibrary. In S. Carver (editor) *Innovations in GIS 5*. London: Taylor and Francis, pp. 59–68.

Goodchild, M.F. (2001) Towards a location theory of distributed computing and e-commerce. In T.R. Leinbach and S.D. Brunn, editors, *Worlds of E-Commerce: Economic, Geographical and Social Dimensions*. New York: Wiley, pp. 67–86.

Gravano, L., K. Chang, H. Garcia-Molina, C. Lagoze, and A. Paepcke (1997) *STARTS: Stanford protocol proposal for Internet retrieval and search*. Digital Library Project, Stanford University. http://www-db.stanford.edu/~gravano/starts.html

Haynes, K.E., and A.S. Fotheringham (1984) *Gravity and Spatial Interaction Models*. Beverly Hills: Sage.

Hill, L. L. (1998). Building georeferenced collections: Gazetteer services. *Taxonomy Authority File Workshop, Washington, D.C., June 22-23*. http://www.alexandria.ucsb.edu/~lhill/Gazetteer_Taxonomy_Presentation/Taxonomy_presentation.html.

Hill, L.L., G. Janée, R. Dolin, J. Frew, and M. Larsgaard (1999) Collection metadata solutions for digital library applications. Alexandria Digital Library, University of California, Santa Barbara.

Janelle, D.G., and D.C. Hodge, editors (2000). *Information, Place, and Cyberspace*. New York: Springer.

Lagoze, C. (1998) *STARTS: Stanford protocol proposal for Internet search and retrieval: reference implementation*. http://www2.cs.cornell.edu/lagoze/starts/starts_reference.html

Mapping Science Committee, National Research Council (1999) *Distributed Geolibraries: Spatial Data Resources*. Washington, DC: National Academy Press (in review).

Montello, D.R., M.F. Goodchild, P. Fohl, and J. Gottsegen (1998) Fuzzy spatial queries in digital spatial data libraries. *Proceedings, FUZZ-IEEE 98, 1998 World Congress on Computational Intelligence, Anchorage, Alaska*.

Murnion, S., and R.G. Healey (1998) Modeling distance decay effects in Web server information flows. *Geographical Analysis* 30(4): 285–303.

Olsson, G. (1965) *Distance and Human Interaction: A Review and Bibliography*. Philadelphia, PA: Regional Science Research Institute.

Wilson, A.G. (1970) *Entropy in Urban and Regional Modelling*. London: Pion.

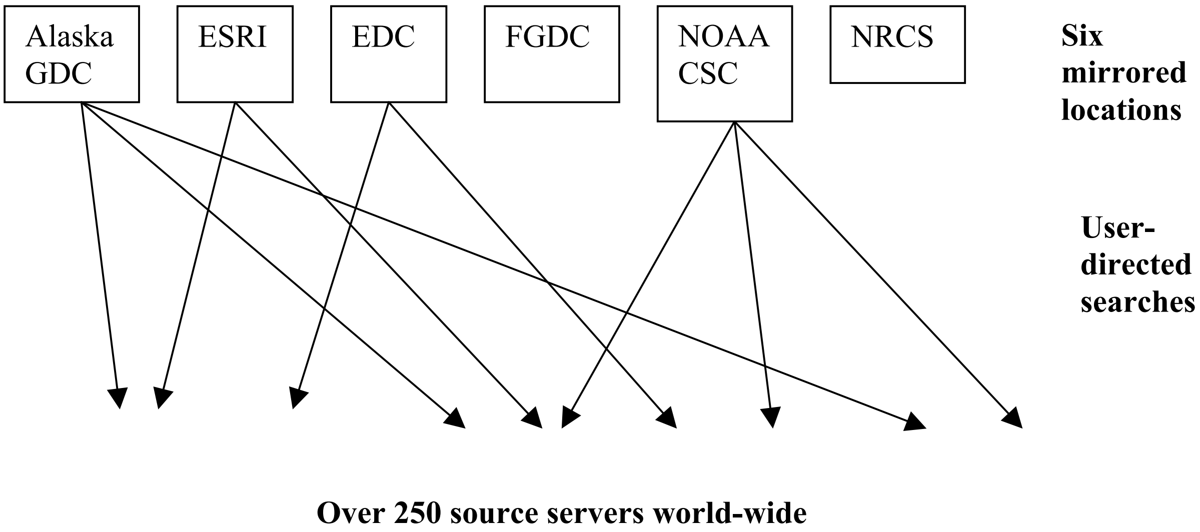

Figure 1: The configuration of the U.S. National Geospatial Data Clearinghouse (in early 2003)

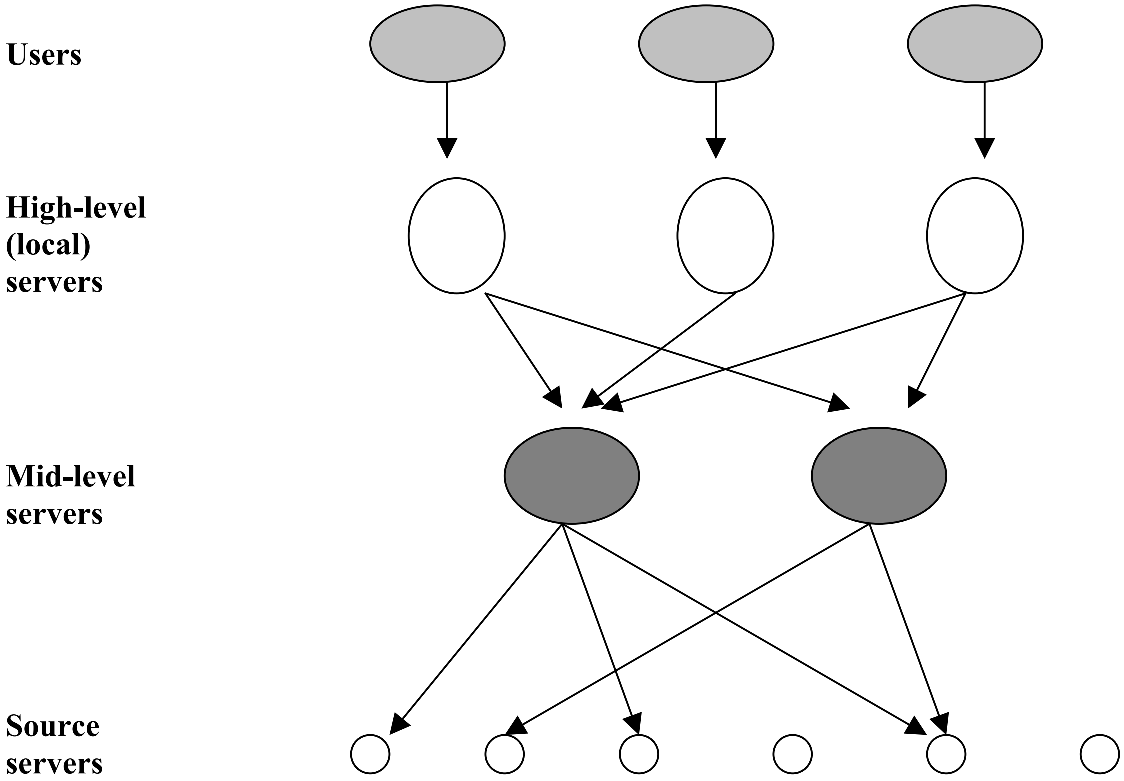

Figure 2: The Pharos configuration, after Dolin *et al.* (1997)

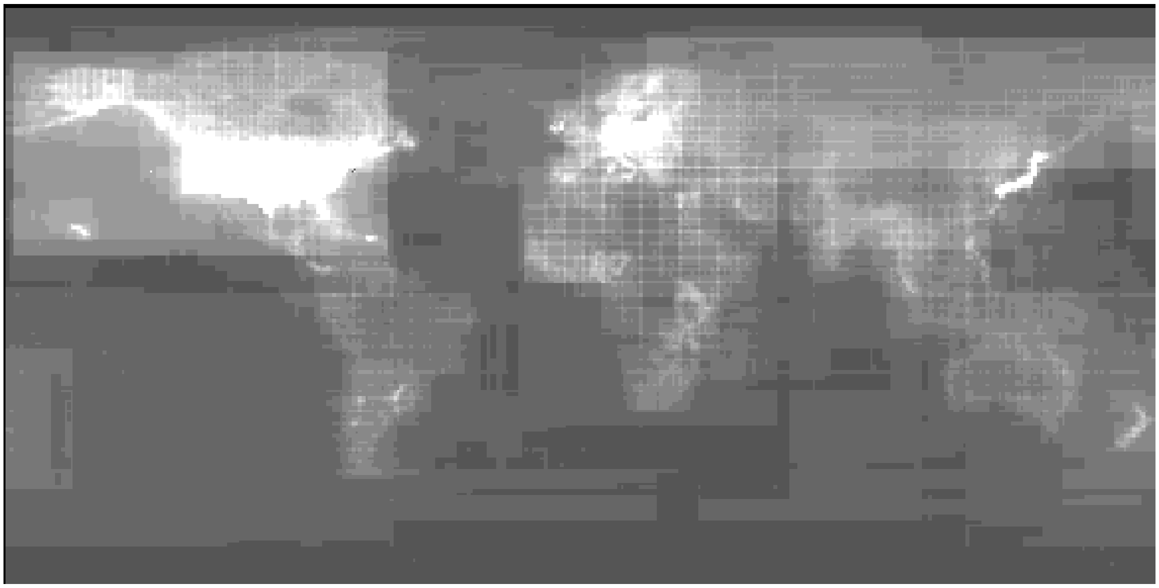

Figure 3: Frequency of GIBOs in the ADL collection by one-degree cell (white indicates the highest frequencies)

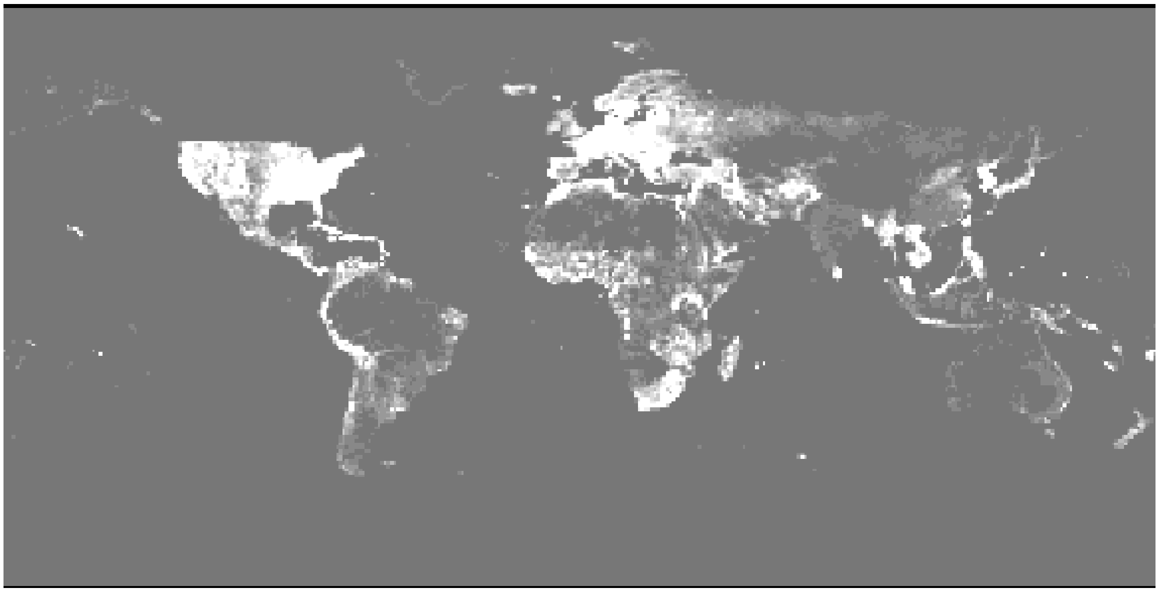

Figure 4: Frequency of entries in the ADL gazetteer by one-degree cell (white indicates highest frequency)