The GeoWeb — A New Paradigm for Finding Data on the Web

Yvan G. Leclerc *Martin Reddy* *Lee Iverson* *Michael Eriksen*

SRI International, Menlo Park, CA 94025, USA.
{leclerc|reddy|leei|eriksen}@ai.sri.com

1. Introduction[1]

The Web has revolutionized the way documents are published, found, and viewed. This revolution has succeeded in part because Web protocols and publishing standards are open, and there exist freely available tools for publishing and reading Web documents. Just as importantly, powerful search engines let users find documents of interest almost instantaneously. Without these search engines, the Web would be almost useless.

Even though search engines can now find almost all Web-accessible documents about a given topic, it is currently impossible to find almost all web-accessible data about a given location. (Examples of such georeferenced data include aerial and satellite images, 3D models of buildings and weather systems, and vacation pictures taken with GPS-enabled cameras.) That's because the location that the data refers to is typically not a part of the data itself. Thus, today's search engine technology is not applicable and the huge amount of georeferenced digital data that is available on the Web today is virtually inaccessible.

In an attempt to make georeferenced data accessible, a number of companies and organizations have created private or governmental databases that hold a small fraction of all georeferenced data. Companies like Mapquest maintain private map databases with associated street addresses and links to businesses like restaurants and shops. Organizations like the Federal Geographic Data Committee (FGDC) maintain databases with government-owned imagery and feature data. These databases hold searchable *metadata*, a summary description of the data that includes geographic location, which can either be searched directly, or via a clearinghouse. However, none of these organizations is either capable of, or willing to, create a database that would allow anybody in the world to publish and search georeferenced metadata. Indeed, this task is so large that no single, regional, organization could do it alone.

What is needed instead is a coordinated global infrastructure with participating organizations from around the world. We call this infrastructure the GeoWeb. We propose to build and maintain this open standards-based infrastructure on a new top-level domain called .geo that will enable anybody to publish and search for all metadata referring to a given area. The infrastructure is based on a hierarchy of servers whose domain names represent geographic areas.

[1] The work performed by SRI International was funded in part by the Defense Advanced Research Projects Agency (DARPA) under contracts MDA972-97C-0037, subcontract 12165SRI of contract no. F19628-95-C-0215), and MDA972-99-C-0011. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Defense Advanced Research Projects Agency, the United States Government, or SRI International.

An example hierarchy, described below, is nominally of the form minutes.degrees.tendegrees.geo. (For convenience, we use ".geo" as the top-level domain name in all of our examples of the GeoWeb hierarchy.)

Using the GeoWeb hierarchy, Web sites or specialized applications will let users specify a search profile (keywords and other pertinent information) and find all georeferenced data satisfying the profile in a given area. Because the data is georeferenced it can be embedded in 2D maps or 3D models of the Earth, letting the users navigate through the map or model, providing a much more natural means of finding data than traditional search engines.

In the remainder of this paper, we will describe the GeoWeb hierarchy and metadata, and some client applications.

2. Technology

The GeoWeb is a distributed searchable database of metadata. It comprises the metadata standard, the distributed searchable database in the form of a DNS hierarchy of servers, and an API library for publishing, searching, and validating/endorsing metadata in the servers.

2.1 GeoWeb Design Principles

We have designed the GeoWeb according to a few basic principles (similar in many ways to the Alexandria Digital Library):

1. **Naive transparency**: We want to make sure that the mapping from categories to metadata content is syntactically and semantically trivial. It should not be a difficult task to put together a sophisticated query involving a combination of criteria for a search.
2. **Latitude/longitude and temporal extent**: Since the basic search criterion is lat/long, elevation, and time, these must be explicit and required fields in the metadata.
3. **Efficient coverage**: In general, search fields should all be of relatively equal relevance. This way, we don't have the overhead of making something explicitly available for search that is only rarely actually used for such.
4. **Abstract query formats**: We want to explicitly avoid exposing any of the internal structure of the database in order to formulate queries. Moreover, we need to allow for the possibility of sophisticated boolean combination of simple queries in order to find exactly the desired set of metadata.
5. **Standard format**: We would like to be able to adopt or at minimum support one of the existing standard metadata formats for external publishing (e.g. Dublin Core, FGDC, ISO/TC211).
6. **Multilingual**: Given that we seek to be globally useful, it is important to respect the need to publish and discover metadata in languages other than English. This demands that we support both multilingual descriptions and cross-linguistic categorizations of objects and services (e.g. the Universal Standard Products and Services Classification (UNSPSC) system).
7. **Privacy and Information Security**: It is vital for many potential uses of the GeoWeb infrastructure to ensure that privacy can be maintained in the face of a publicly accessible

searchable infrastructure. We want to have a general framework for restricting access to links, disallowing "tracking" of dynamic metadata items and allowing third parties to validate or endorse metadata entries.

2.2 GeoWeb Metadata

As we have developed the concepts that lead to the GeoWeb, it has become clear that we have a somewhat different view of what may be considered georeferenced information and thus envision a very different kind of metadata than is traditional. Traditional metadata (e.g. Dublin Core, FGDC or ISO/TC211) is very much a description of a particular, single data item or service (e.g., a single web site, web page, aerial image, or photograph). The searchable metadata we have been developing is different from this in a fundamental way: we wish to provide a semantic and geographical description of physical or conceptual objects that may have a number of different manifestations as data or services.

For example, a physical object such as a restaurant may have a metadata entry keyed to its location that contains a description and links for several distinct data elements (e.g. a home page, a menu, and a VRML model of the building) and services (e.g. a Voice over IP address, an online takeout order form, etc.). A user should be able to search for "a South Indian restaurant with an online menu and photographs within walking distance of my hotel." While it makes a great deal of sense to provide a semantic nexus for information at this level, this does not correspond to a traditional view of geographic metadata.

While this may seem discouraging, the semantic break here is not actually that great. It should be clear that traditional geographic metadata is actually encompassed by this broader conceptualization. A publisher could consider a single piece of data to be the appropriate conceptual object to be described by a metadata entry. It should be up to the metadata publisher to choose which level of object description is most appropriate for representing their information within the index. Given that the choice to register a metadata entry is a very simple form of advertising, this flexibility can be considered to be one of the advantages of the opt-in registration strategy we have adopted.

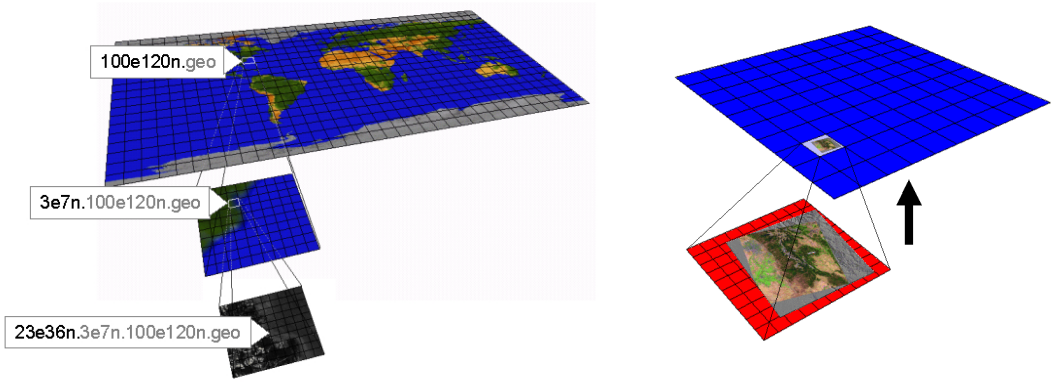
2.3 The GeoWeb DNS Hierarchy of Servers

Metadata in the GeoWeb is distributed across a number of servers for scalability. The servers have distinct Internet Domain Names that identify regions on the Earth bounded by latitude and longitude. Such a region is called a *cell*, and the domain name that identifies it is called a *geographic domain name*. (See Figure 1(a)). Following are example applications of geographic domain names using the minutes.degrees.tendegrees.geo name schema. (This schema is used here for simplicity. The schema we currently favor is actually of the form tenminutes,degrees.tendegrees.ninetydegrees.geo.)

- The geographic domain name *20e30n.geo* identifies the 10-degree x 10-degree cell whose southwest corner is located at 20 degrees east, 30 degrees north.
- The geographic domain name *2e4n.10e50n.geo* identifies the 1-degree x 1-degree cell whose southwest corner is located at 12 degrees east, 54 degrees north.

- The geographic domain name *11e21n.3e7n.30e10n.geo* identifies the 1-minute x 1-minute cell whose southwest corner is located at 33 degrees, 11 minutes east and 17 degrees, 21 minutes north.

Metadata is placed in the hierarchy according to the number of cells that overlap the geographic coverage of the data it corresponds to (see Figure 1(b)). Specifically, we use the level with the smallest cells such that no more than four cells overlap the geographic coverage. At this chosen level, the metadata is stored in all the overlapping cells.



(a) (b)

Figure 1. (a) Illustration of the GeoWeb hierarchy showing example server names, (b) demonstration of the procedure to locate the appropriate cell server(s) to store a new item in.

This naming convention and metadata placement strategy allows clients to determine which host(s) to query for metadata referring to a given area, thereby distributing the load over many servers with no single point of failure or congestion. This distribution and query method also reduces the storage and bandwidth requirements of any given server, making it possible for thousands of high-speed visualization systems to be simultaneously searching for and retrieving metadata and data.

The use of DNS also allows metadata to be transparently moved to new hosts and subdomains as needed. For example, initially all metadata may be physically hosted on a single computer, with all .geo domain names aliased to that machine. But over time, as areas fill with metadata (such as large urban areas), their corresponding subdomain and metadata can be transparently transferred to new computers. These computers may be maintained by other organizations who may, when necessary, transfer data to new computers (and new organizations) deeper in the hierarchy.

2.4 Dynamic GeoWeb Content

Many objects that might gain some advantage by being registered in the GeoWeb do not have fixed positions at all, and it is their current position (and potentially history and planned future of movement) that should be reflected in the GeoWeb. These *dynamic* objects require a somewhat

different interface to maintain their positions in the GeoWeb Hierarchy and their associated geographic domain names. We have designed a facility whereby an intermediate directory server can manage the metadata for such objects based on a globally unique object identifier instead of geolocation, This intermediate server then translates updated locations to the appropriate geographic domain name and updates the metadata in the GeoWeb Hierarchy, taking care to remove metadata entries when they dynamically move out of particular cells. These intermediate servers may be either public or private (depending on the need to advertise or hide the object-ID to metadata mapping) and may offer other associated services. By structuring the solution this way, we retain all of the distribution and localized discovery advantages of the GeoWeb architecture while allowing dynamic objects to efficiently update their locations.

For example, an aircraft equipped with GPS receivers and secure wireless Internet access could periodically contact the dynamic directory server to update its current location, velocity, and other attributes. This information would then be used to update the geographic coordinates (and other attributes) in the corresponding metadata record in the GeoWeb hierarchy, and, when necessary, transfer the metadata record to a new cell. This dynamic metadata can then be used for many purposes, including air traffic analysis and providing global near-real-time maps of air traffic. Of course, the aircraft could in turn use the GeoWeb to download or access localized flight/airspace support information.

In the case of weather data, the dynamic directory server might be maintained by the U.S. National Weather Service itself, since its mandate includes the maintenance of real-time, public weather maps and storm information. As in the example above, the dynamic directory server would periodically update the corresponding metadata record and, when necessary, transfer it to a new cell.

Should the resource and network loads on these dynamic directory servers become too great for traditional single-point database methods, it would be relatively straightforward to take advantage of the strategy used by the location-based GeoWeb hierarchy and create a distributed DNS hierarchy to represent an object-identity index. For example, airplanes already have a registration ID that could be translated (and possibly encrypted for privacy) to a DNS name/id pair such as *us.air.geo/AC7321*. This server/id pair could then be used with exactly the same protocol as is used to update GeoWeb registry cells to update a dynamic directory server which then automatically updates the associated entry in the location-based GeoWeb hierarchy. In database terms, we can provide a second primary index into the global metadata database using exactly the same update protocols. We reiterate that these secondary indices may be either private or public, depending on data privacy concerns in the particular domains implemented.

2.5 Validation of Metadata and Data.

To ensure the accuracy and validity of certain classes of data for certain purposes (such as elevation data used for city planning purposes), the metadata record has a field for one or more optional validation certificates issued by validation organizations. The digitally signed certificate will certify that one or more elements of the metadata/data meet certain qualifications, such as accuracy or completeness, or compliance with local, national, or international standards or conventions. While essential for validation purposes, this facility also has a number of other uses,

such as for reviews. For example, an organization such as the American Automobile Association could annotate hotel and restaurant listings with references to their own independent descriptions of the facilities and quality of service that these businesses provide. Clearly this facility has further implications for security and privacy than those already outlined above.

3. Advantages

The GeoWeb infrastructure offers a number of advantages over today's state-of-the-art technology. It was specifically designed as an extremely scalable, open, and global geographical index to the Web. Below are a few of the key benefits of this technology.

- **Open opt-in scheme.** No one company or institution can hope to index all geographically related information for the whole planet. Therefore, the GeoWeb is an opt-in scheme where any user can register their own data, be it information about their bed and breakfast, or pictures of their trip to the Big Island.
- **Integrate disparate data sources.** Instead of being restricted to one company's view of the world, the GeoWeb allows users to discover and integrate disparate sources of information for improved decision making.
- **Massively scalable architecture.** The entire GeoWeb index could be stored on one server, but as demands and database size grow individual cell servers can be split off on to additional physical machines. To illustrate the inherent scalability, there are over 60,000,000 1-minute servers possible over the land regions of the earth.
- **No one-server bottleneck.** Most contemporary search and indexing interfaces require going through a single point-of-failure server. In the GeoWeb, the client works out which cell server to contact based on the geographic extent of the query.
- **Transparent load balancing.** Using the DNS hierarchy enables transparent load balancing by moving data to new servers and organizations as needed. The IP addresses change and the bandwidth improves, but the domain names remain the same.
- **Index any content.** The GeoWeb can index any content, not just HTML. Movies, sound files, text, 3D models, terrain, etc. can all be indexed. Content such as HTML with geographic meta tags allows an easy way to register that content in the GeoWeb.
- **Built-in security.** Security is imperative to ensure that private data remains private, while still providing easy access to public information. In addition, there is a scheme to allow appropriate organizations to validate and/or endorse data..
- **Support high bandwidth demands.** Massive distribution of the geographic database enables many clients around the world to retrieve data simultaneously and at high speed.
- **Reduced server costs.** As the GeoWeb infrastructure is distributed over many cell servers, there is no need for a single massive server with huge storage requirements. The size and speed needed per server is therefore reduced, for example, divided by 10,000 or more for 10-minute servers.
- **Local control.** Companies, institutions, or local governments can manage the cell servers for their region of the world, and also host the cell servers in their geographic region.

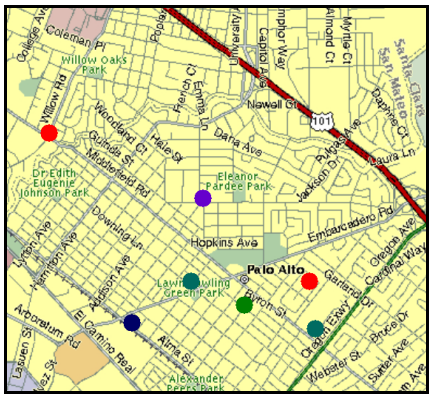
4. GeoWeb Clients

Naturally, users will not normally browse the GeoWeb by typing in domain name addresses such as *11e21n.3e7n.30e10n.geo*. The benefit of the GeoWeb is that a client application or portal can trivially work out the server to contact for any given region of the earth. A range of possible GeoWeb clients can be envisioned. The following sections introduce a few of the interfaces that we have proposed or are already developing.

4.1 Text-based

The most basic kind of interface would be similar to the types of search engines that we use today. However, in addition to entering a search criterion, you would also enter a location for your search. This could be a latitude/longitude coordinate, or more intuitively, a street address or a feature name that can be translated to a geodetic coordinate using a gazetteer service. The results could be presented as a simple list of links (ordered by distance), as is the case for most search engines today. This type of interface has the advantage that it scales down well to current handheld devices such as a PalmPilot™ or iPad™ (Figure 2(a)).

(a)



(b)

Figure 2. Potential GeoWeb clients, (a) a handheld interface, (b) a map-based interface with hyperlinked icons.

4.2 Map-based

There are already various mapping sites on the Web that will generate a 2D map for a region of interest, e.g., MapBlast!, Yahoo! Maps, MapQuest, Maps.com, etc. It is easy to imagine a GeoWeb client that lets users search over a particular geographic region and return the results as a map image with various icons overlaid (see Figure 2(b)). These icons could be hyperlinked so that clicking over them would take you to more information about that item, for example, clicking over the icon for a restaurant might take you to their menu or on-line reservation page.

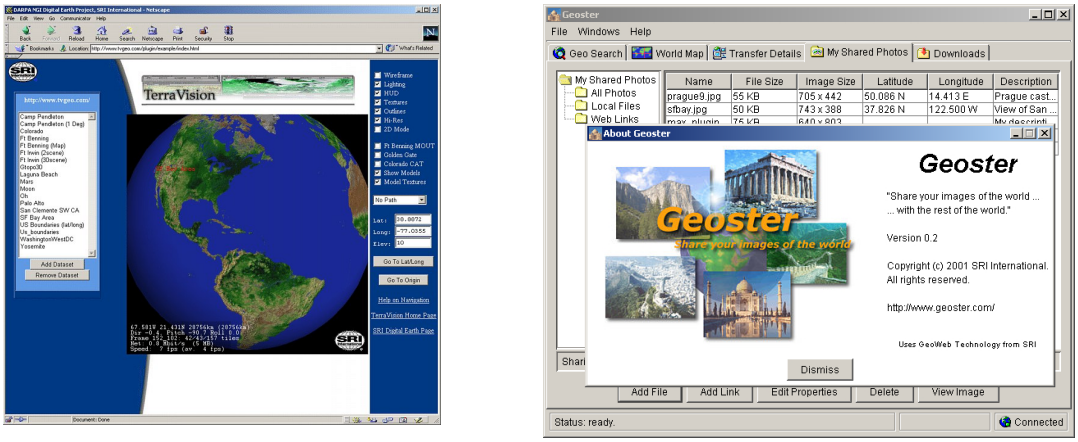
4.3 TerraVision

TerraVision™ is a distributed, interactive terrain visualization system developed by SRI International. It allows users to navigate, in real time, through a 3-D graphical representation of a real landscape created from elevation data and aerial images of that landscape. TerraVision can browse huge datasets, in the order of terabytes, and these data can be distributed over multiple servers across the Web. 3-D GeoVRML models can be overlaid on the terrain, such as building models, atmospheric simulations, icons, place names, etc. We have instrumented TerraVision to become a GeoWeb client so that, as the user flies around the world, the system is constantly

querying the GeoWeb index to discover all relevant information about the place being viewed. This information is then streamed automatically to the user's display for them to browse visually. TerraVision is freely available for Unix and Windows platforms from http://www.tvgeo.com/, including an ActiveX component and Netscape plug-in under Windows (see Figure 3(a)).

4.4 Geoster

The popularity of file sharing protocols and applications has been amply demonstrated with the enormous usage of services like Napster and Gnutella. We at SRI are developing a file sharing application along the lines of Napster, but using the GeoWeb as the underlying indexing and search infrastructure. Geoster (pronounced *joy-ster*) is an application built for sharing photos over the web using the GeoWeb (see Figure 3(b)). It lets anyone to index their photos from around the world and to share these with the rest of the world. This would allow you to do searches such as, show me all photos that have been taken of the Great Wall of China. Geoster is freely available from http://www.geoster.com/ and is written in Java and hence available across multiple platforms. We believe that this type of client could produce a large interest in the GeoWeb and generate a substantial amount of GeoWeb registrations.



(a) (b)

Figure 3. Sample GeoWeb clients, (a) TerraVision plug-in, (b) Geoster photo-sharing client.

5. Conclusion

As a fundamental new service on top of the Web, we believe that the GeoWeb will provide a unique paradigm for harnessing the power of the Internet. It will be based on the way human beings perceive and comprehend their world: geospatially, in three dimensions, and over time. The GeoWeb will enable Internet users to navigate, access, and visualize georeferenced data as they would in a physical world, but without the barriers imposed by space and time in the physical world. It makes the world knowable as never before.