
**The New York State
Spatial Data Clearinghouse
Technical Report**

CTG.GIS-003



**Center for Technology in Government
University at Albany / SUNY**

ã 1995 Center for Technology in Government
The Center grants permission to reprint this document provided that it is printed in its entirety

Center for Technology in Government Project

In 1994 the Department of Environmental Conservation (DEC) proposed a project to the Center for Technology in Government (CTG) to help create a framework within which New York State organizations could pursue their individual business objectives within the context of shared spatial information resources. The proposal centered on the idea of a data “cooperative” where many organizations would work together to share data and expertise for individual and mutual needs. A readily accessible spatial data clearinghouse was a key feature of the cooperative concept. The project was selected from a field of 19 proposals. Efforts to define the focus and scope of the project began in December 1994 with a small team of CTG and DEC staff. The project team quickly expanded to include involvement by a number of state, local, and private organizations. The project team included:

- Department of Environmental Conservation IRM and GIS staff
- Faculty experts from the University at Albany
- Faculty and staff from the University at Buffalo’s National Center for Geographic Information and Analysis
- The State Archives and Records Administration staff
- NYS Forum for Information Resource Management staff
- Erie County Water Authority staff
- CTG professional staff and graduate students

The project team also included 10 corporate partners. Applied GIS, Inc. was instrumental in the design of the NYS Spatial Data Clearinghouse. Blue Moon Training Systems, Inc. provided the CTG project staff with training in Geographic Information Systems. Xyplex, Aule-Tek, Inc., and Sun Microsystems provided CTG with support for remote access to the Clearinghouse. Harlan Wallach, a graphic artist, designed the logo for the Clearinghouse. Inteligis Corporation, a not-for-profit GIS consulting firm near Buffalo, provided support for some of the research efforts undertaken in the project. Documentation Strategies, Inc. together with Full-Circle Communications provided CTG with technical writing expertise for the Clearinghouse. Digital Equipment Corporation provided consulting services for some of the project activities.

Project participants came from all levels of government, the private sector, the academic community, utilities, and the not-for-profit sector. The project research activities involved approximately 450 participants from across the state and many additional participants from the national GIS community. Presentations and articles in regional user group newsletters were used to inform the GIS community about the project activities and to invite participation, reaching about 3,500 individuals.

Project Objectives

The Center worked with the Department of Environmental Conservation to improve the ability of any individual or organization to identify the existence of relevant data sets, to enhance understanding of the value of geographic information systems, and to explore the need for coordination of efforts to use GIS and spatial data in New York State. The project team identified a set of deliverables that reflected the proposal submitted by DEC, and which, to the extent possible, also addressed some of the charges to the NYS Temporary GIS Council which was getting underway at the same time.

The project activities were focused in two areas. The first was the development of a prototype designed to demonstrate the efficacy of an on-line clearinghouse of metadata and spatial data sets. The Clearinghouse would be available to public, private, academic, and non-profit users as a mechanism to share data. The federal metadata standard was adopted for use in the prototype Clearinghouse. In a parallel project activity, the federal standard for metadata was further analyzed for its usability to support data sharing in New York State.

The second focus was to review the literature and work with the GIS community to gather data on the value of GIS as a decision making tool, to identify effective approaches to assessing costs and benefits, to identify barriers to sharing and coordination of GIS activities, and to gather information and recommendations from the community regarding the future coordination of GIS in New York State.

Within this larger framework, the project team pursued three specific objectives:

1. Demonstrate the value of GIS by examining exemplary applications and existing evaluation approaches.
2. Identify barriers to sharing spatial data and explore potential solutions for overcoming those barriers.
3. Investigate some practical tools to support GIS coordination in New York State.

This paper reports on the project efforts associated with the development of a web-based prototype data repository as a mechanism to support GIS coordination in New York State. An overview of the larger environment within which the NYS Clearinghouse exists, the functionality of the Clearinghouse and the underlying technical structure are presented. Implementation alternatives and recommendations for implementation are also presented. (Please see Appendix A for a related products list.)

Initiatives at the Federal Level

The federal government has created a comprehensive national initiative focused on the value of spatial data. The federal government has begun to address GIS issues with the

establishment of the National Spatial Data Infrastructure (NSDI). According to FGDC, the NSDI is “a set of policies, standards, materials, technologies, people and procedures, as well as spatial data, that provide a foundation for more efficient collection, management, and use of data. The goal is better access to higher quality spatial data at lower costs to all. The NSDI requires cooperation and interaction among various levels of government, the private sector, and academia.” The major components of the NSDI are:

- Standards to facilitate data collection, documentation, access, and transfer
- A basic framework of digital spatial data that meets the minimum needs of large numbers of data users over any given geographic area
- A clearinghouse to serve, search, query, find, access, and use spatial data
- Education and training in the collection, management, and use of spatial data.

While the NSDI is managed at the federal level under the leadership of the Federal Geographic Data Committee (FGDC), many state and local governments, as well as academic and private sector organizations, have joined the effort to promote better access to spatial information, to increase communication and cooperation within the GIS community, and to eliminate costly data redundancy. The efforts of all of these organizations together form the NSDI.

The National Spatial Data Clearinghouse (NSDC), an Internet-based tool to facilitate search and retrieval of spatial data sets, is a key part of the NSDI. A number of federal agencies and state governments have built spatial data clearinghouses on the Internet that can be accessed from the NSDI home page. The URL for the NSDI home page is <http://fgdc.er.usgs.gov>

The Prototype New York State Spatial Data Clearinghouse

The NYS Spatial Data Clearinghouse was developed in concert with the Federal initiatives to facilitate the exchange of spatial data among members of the NYS GIS community. The National Spatial Data Clearinghouse (<http://fgdc.er.usgs.gov/fgdc.html>), available on the Internet, is designed to increase the value of spatial data through sharing. As part of this national effort, the prototype NYS Clearinghouse provides a mechanism for potential users of NYS spatial data to determine whether data sets they need are already available or under development. This means of improving access to and sharing of spatial data has the potential to lower the cost and greatly increase the use of these data throughout New York.

The NYS Clearinghouse is unique among state Clearinghouses for two primary reasons. First, it is a statewide resource available to support the sharing of spatial data for all sectors. State agencies, local governments, non-profits, academia, utilities, and the private sector may all use the Clearinghouse as a mechanism for sharing their spatial data as well as for identifying useful spatial data. Second, the NYS

Clearinghouse and those of Wisconsin and Montana are the only state clearinghouses which offer spatially searchable standard metadata.

The primary purpose of the Clearinghouse is to allow producers of geographic data to describe what data sets they have available and to allow users of Geographic Information Systems to find the data sets they need. Once the appropriate data set is located, the system provides information on how to obtain the data files, including an option for immediate online transfer of the files using ftp.

Users who access the Clearinghouse may search for available data and review detailed descriptions of the data. Once a data set of interest is identified, information is provided on how to obtain the data files. For some data sets, an option for immediate on-line transfer of the files is available using standard file transfer protocol (ftp).

The URL for the prototype NYS Spatial Data Clearinghouse is:
<http://www.ctg.albany.edu/gisny.html>. Figure 1 shows the World Wide Web home page for the Clearinghouse.

Describing Data Sets with Standard Metadata

Standard metadata (information describing data) is an essential prerequisite for effective information sharing. Contributors to the Clearinghouse described their data sets using the FGDC Content Standard for Digital Geospatial Metadata (<http://fgdc.er.usgs.gov/metaover2.html>). The FGDC standard specifies what should be contained in a metadata record for spatial data sets. It includes such information as who produced the data, the geographic area covered, the data set category or theme, scale, accuracy information, and instructions for how to obtain the data set. While the FGDC standard may be further refined for New York State, it proved to be adequate in demonstrating the functionality of the GIS prototype. Results from the prototype also demonstrate that the standard could be used as-is in a production environment until further refinement takes place.

Standard metadata is critical to collaboration and exchange of data sets. A standard data description offers these necessary components for a successful spatial data clearinghouse:

- **Standard metadata provide a common language for GIS users to describe data sets.** Since one goal is to facilitate the exchange of data sets among a large number of independent organizations, a common language for describing the data sets is needed.
- **Standard metadata help ensure that the data sets will be described thoroughly.** The metadata template identifies mandatory and suggested fields; consequently, it serves as a guideline to those describing the data

sets. The metadata can serve as an important tool for an organization's internal documentation as well as for data set exchange.

- **Standard metadata offer the opportunity for automation.** Once the metadata are standardized, software can be developed for creating, collecting, and searching the metadata on the Internet.

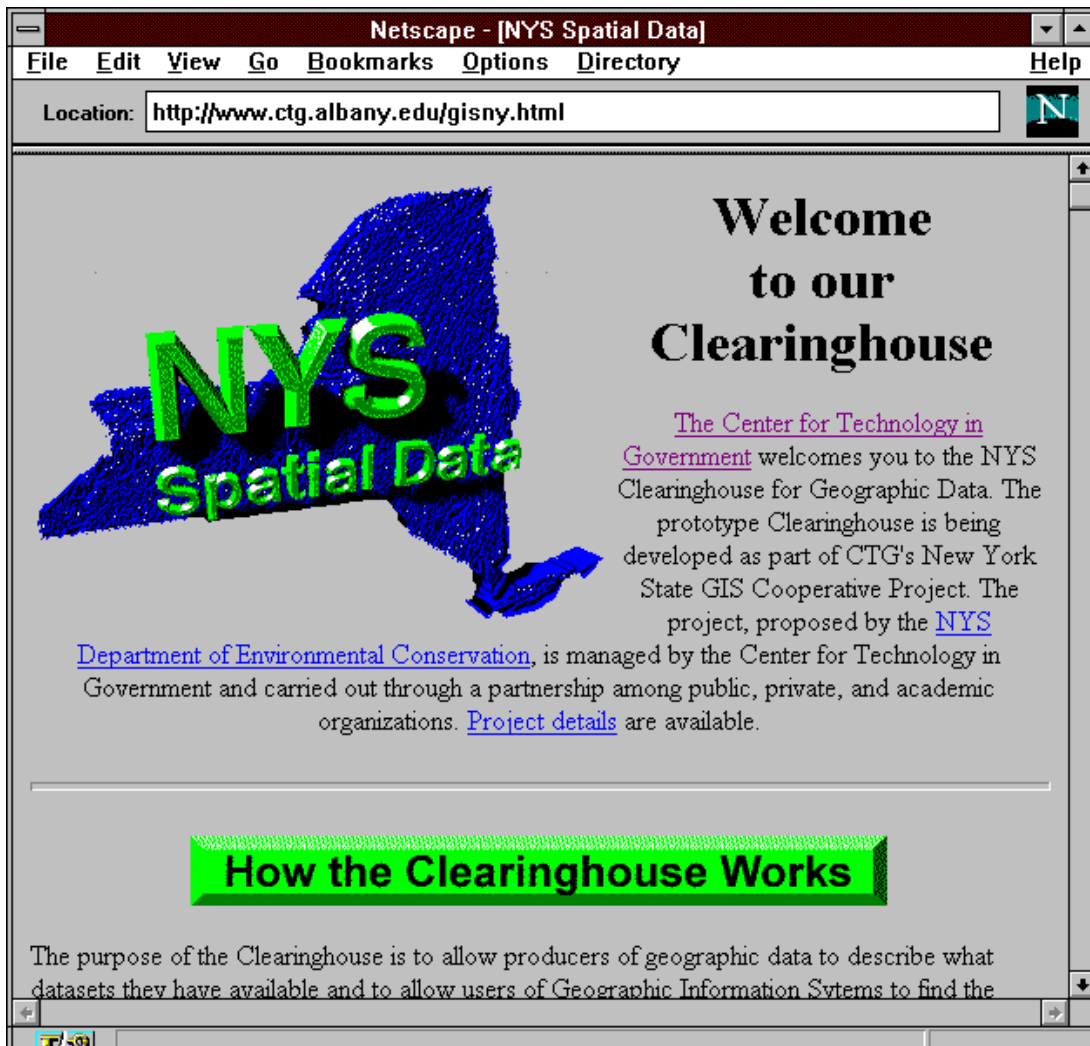


Figure 1 - NYS Spatial Data Clearinghouse Home Page

Search and Retrieval

The Clearinghouse offers two methods for data set identification. The first method is a directory which lists all available data sets, organized by category and distributor.

Users can browse through the directory to find the data sets they need. The second method uses a search form (Figure 2) to enter the specific criteria required.

Both the directory list and the search results list contain links to the full metadata record. When a data set is of interest, the full metadata document can be reviewed in order to obtain a fuller understanding of its properties. In some cases, the metadata document also contains an image which reflects the geographic region and features that the data set encompasses. The distribution section of the metadata contains instructions for obtaining the data set. Those instructions, supplied by the metadata provider, may include on-line file transfers, electronic order forms or instructions for ordering by phone or mail.

The Search Form

The search form is an HTML forms document which allows users to enter their criteria for locating data sets. Three options are available for searching:

- **Geographic Area** (mandatory): The geographic area can be selected by clicking on an area identified in the displayed list or by entering the latitude and longitude coordinates for the bounding rectangle of the desired area. The search will find metadata records whose bounding coordinates overlap any part of the area chosen. When the user selects an area using the displayed list (instead of supplying actual coordinates), the software automatically selects corresponding coordinates for the area and then conducts the search.
- **Data set Category or Theme** (optional): To narrow the search further, the user can select one or more of eighteen categories. The categories represent broad data set themes, such as Cadastral, Demographics, Environment, or Infrastructure. A Geo Thesaurus (Figure 3) has been constructed to provide further help in choosing the appropriate category. When the Thesaurus option is clicked, the screen displays a table of common terms and identifies which categories may be useful for finding related information. For example, the term 'tax map' points to Cadastral, while 'roads' points to Infrastructure.
- **General Query** (optional): Words listed in this input box are matched against the entire metadata document. Desired words can simply be listed in the input box, separated by spaces, or more complex Boolean searches can be constructed using AND, OR, and parentheses. Right-hand truncation is also permissible with the use of an asterisk, so that sch* will find all documents which contain words beginning with "sch."

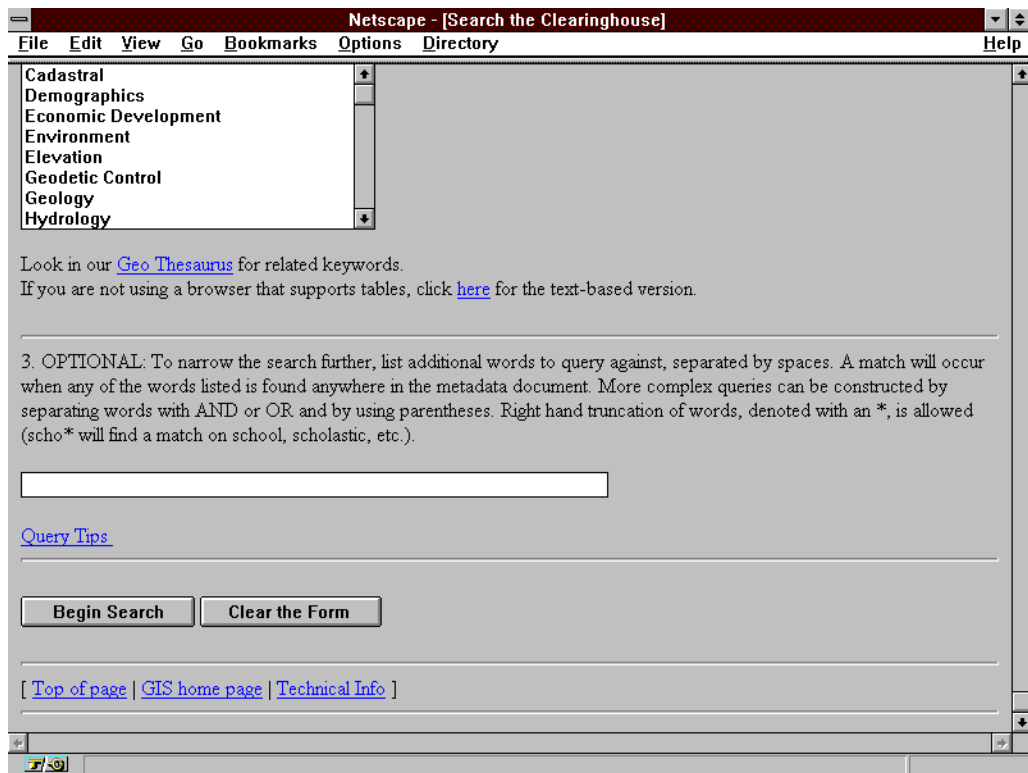
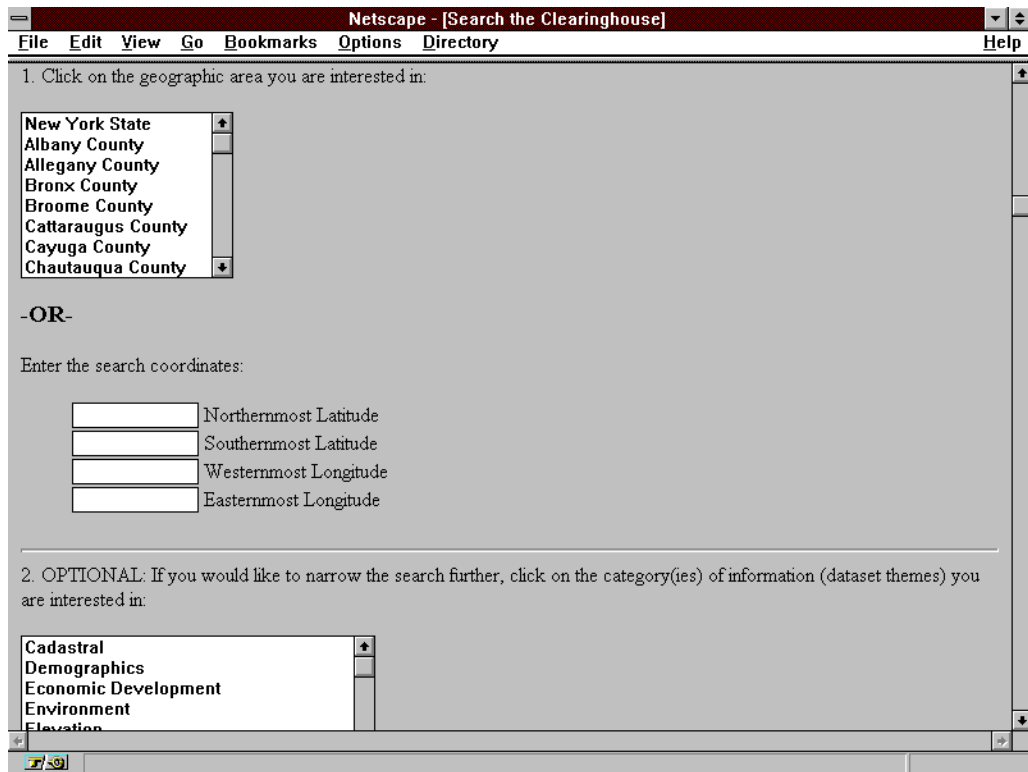


Figure 2 - Search Form

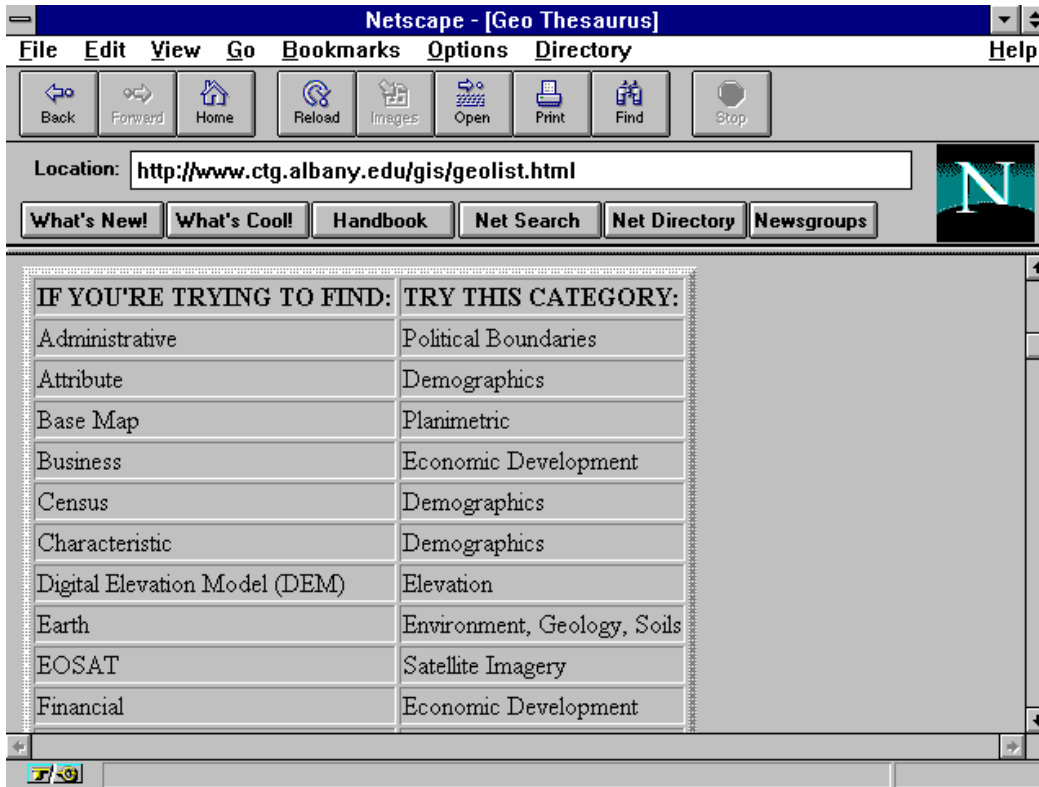


Figure 3 - Geo Thesaurus

Clearinghouse Implementation & Reaction

The NYS Clearinghouse went on-line in August, 1995, with eight organizations providing metadata for nearly fifty data sets. The Federal Geographic Data Committee created a link to the NY site from the National Spatial Data Clearinghouse under its State and University listings. The NYS Clearinghouse was also registered with general Internet search sites such as Yahoo and Lycos. By late December, over 2,500 visits were made to the Clearinghouse and that number was growing at a rate of about 50 visits per business day.

User reaction has been uniformly positive. An on-line survey gathering both general user information and specific feedback on the Clearinghouse was included in the prototype. The survey questions were designed to collect information about users and about their reaction to the Clearinghouse. User questions sought information about each visitor's level of Internet and GIS experience, search methods, and primary interest in visiting the Clearinghouse. Other questions asked how the search form, instructions, and results listings could be improved and what additional information could be included in the Clearinghouse. Few suggestions were made for improvement and those made were primarily confined to a desire for more universal use of existing

Clearinghouse features, such as automatic downloading of files through ftp and use of geographic map images to describe the data sets.

Experience with the Clearinghouse has motivated some new efforts to provide metadata and spatial data. The Department of Environmental Conservation was an active participant in the design and implementation of the Clearinghouse and is now in the process of developing its own site for serving spatial data. Other positive reactions came from local governments. After attending a demonstration of the Clearinghouse, the Orange County Water Authority provided metadata for six data sets. Orange County Water Authority staff then demonstrated the Clearinghouse to other county officials to build support for providing full access to the County's data resources via a clearinghouse mechanism. As a result, Orange County is implementing a local server connected to the NYS Spatial Data Clearinghouse to provide no-cost on-line access to metadata and spatial data sets. Rockland County is considering a similar program.

Despite these early successes, much remains to be learned about repository management and about willingness and ability to use a web-based repository to support data sharing. As the prototype Clearinghouse contains less than fifty metadata descriptions, the full range and volume of use could not be tested. These and other user acceptance and performance issues will need further analysis as the Clearinghouse develops and grows.

How the Clearinghouse was Constructed

The metadata records were collected using a simple metadata template that the contributors could update using any word processor. The completed records were placed in a directory on the CTG server (Digital Equipment Corporation's DECstation 5000 with a RISC processor running ULTRIX 4.2) and were indexed using a software tool called freeWAIS-sf. FreeWAIS-sf was developed by Ulrich Pfeifer at the University of Dortmund, Germany, and is freely available to the public (<http://ls6-www.informatik.uni-dortmund.de/freeWAIS-sf/freeWAIS-sf.html>). FreeWais-sf is an extension of freeWAIS, software provided by the Clearinghouse for Networked Information Discovery and Retrieval (<http://cnidr.org/welcome.html>). FreeWAIS-sf provides two general types of indexing. Global indexing allows every word in the document to be indexed so that the entire text can be searched for a given word or query string. Local indexing allows specific fields to be defined within the document for more precise searching. Both of these methods were employed by the Clearinghouse as our search form allows queries against specific fields as well as against the entire metadata document.

The Common Gateway Interface

The search form interfaces with the http (web) server (NCSA httpd version 1.3) through the use of a common gateway interface (CGI). CGIs allow HTML forms to

pass information from the form to the server and to indicate which program on the server should be executed when the data transmission takes place. This mechanism allows the search criteria to be sent to the server and allows programs to be executed which query the WAIS index for matching documents. The CGI was built using two perl script programs as a foundation: Doug Nebert's spquery.pl (<ftp://h2o.er.usgs.gov/wais/www/spquery.pl>) and Jonny Goldman's fwais.pl (<ftp://h2o.er.usgs.gov/wais/www/fwais.pl>). The programs were modified to suit specific needs of the Clearinghouse. Complete details on the construction of the search mechanism for the Clearinghouse, including sample documents and the program code, is available on the NYS Clearinghouse under URL <http://www.ctg.albany.edu/gis/srchdoc.html>

Other Options for Constructing the Clearinghouse

Through the use of common gateway interfaces, the same forms and functionality can be preserved on the client side of an Internet application, while a variety of options can be employed on the back end for carrying out the search and retrieval. The method employed in the prototype assumes that the metadata will be defined in an HTML or text document, with each field preceded by its tag (field name). FreeWais-sf is able to identify specific fields by recognizing these tags.

A number of commercial products are available to perform the document searches conducted at the server. These alternatives to freeWAIS-sf are primarily relational databases which allow you to store metadata and then search the database using a standard query language.

The advantages of each are outlined below.

Advantages offered by freeWAIS-sf:

- Incorporates the Z39.50 standard (<http://www.research.att.com/~wald/z3950.html>) which is widely accepted by the library community and mandated by the Government Information Locator Service (http://vinca.cnidr.org/protocols/z3950/iitf_report.html) and which provides for a common protocol for searching the databases of multiple servers;
- Allows for both field searching and total document searching. Most relational databases allow only field searching, although there are new products emerging which will allow full text searching within a relational database;
- Is freely available;
- Is easy to learn and implement;
- Can search records in many different formats, such as text, HTML, gif, etc.

Advantages offered by a Relational Database:

- Allows the metadata to be constructed using a relational model rather than the current hierarchical model found in the FGDC standard; this model may more

accurately reflect information about the underlying data sets and will allow common sections of the metadata to be created only once for multiple data sets, thus reducing the effort required for metadata creation and maintenance;

- Does not require that field tags be replicated in each document;
- Allows use of the Standard Query Language (SQL), a widely accepted search language with more sophisticated options than that offered by Boolean searching alone;
- Has attracted greater commercial investment and therefore users can expect better documentation, more sophisticated tools, easier interfaces, and a variety of support options.

Isite

The Clearinghouse for Network Information Discovery and Retrieval (CNIDR) has developed software called Isite (<http://vinca.cnidr.org/software/Isite/Isite.html>), designed to incorporate some of the benefits offered by both freeWAIS-sf and relational databases. Isite uses the Z39.50 protocol for communications between the client and server. On the server side, it includes its own text indexer and search engine. Alternatively, it includes an application program interface (API) which can be used to perform commercial database searches.

Other Uses of the Clearinghouse

The Clearinghouse could serve as a "one-stop shopping" center for GIS-related information. Specifically, in addition to providing information on existing spatial data, it could be used to

- Identify the data set needs of organizations looking for collaborators
- Provide information concerning GIS user groups
- Announce upcoming events
- Disseminate technical papers
- Disseminate vendor information
- Provide links to other Clearinghouses

Recommendations for Implementation

The prototype Clearinghouse is a central site for storing metadata and some spatial data sets. A centrally managed site can offer savings, but we also believe that data management must remain in the hands of the data owners. Although it may be appropriate for a central site to contain metadata, the data sets themselves do not need to be stored at the same site. The size and quantity of the data sets may preclude central site storage. Experience with the prototype NYS Spatial Data Clearinghouse yields the following recommendations.

- Fully automate as many aspects of the process as practical including metadata creation, editing, and collection; updating of all Clearinghouse web pages as metadata is added or changed; and data set transfer through the use of ftp and on-line order forms.
- Centralize the home page and technical support for the Clearinghouse. The Clearinghouse application and its technical administration can be managed most efficiently from a single site, avoiding the need for duplicative development and maintenance by multiple organizations. A central site also ensures a uniform presentation of the metadata and provides a principal forum for information exchange.
- Data set owners, who know the data best, should create and manage their own metadata. Metadata creation, update, and deletion needs to be the responsibility of each independent organization. The development of automated tools will make the independent management of metadata possible.
- Organizations should house their own data sets and create their own ftp sites. The metadata can contain a link to the corresponding ftp site so that a simple click of the transfer option initiates the downloading of the file. Because the data location makes no difference to the user, several organizations can work together to form a cooperative ftp site.
- Clearinghouse efforts should be leveraged by incorporating additional information and functionality for the GIS community. Making the site a “one-stop shopping” center for GIS-related activities in NY State will improve upon the utility of the Clearinghouse, increase visits to the site, and ensure more widespread participation in its ongoing development.

Summary

Information technologists can apply the environmentalists’ concept of “reduce, reuse, and recycle” to improve on the management of data. Members of this project sought to maximize the benefits of developing spatial data by reducing redundant data creation efforts, re-using existing expensive data sets, and recycling spatial data information.

The recognition that organizational units can benefit from cooperative data development and exchange is part of an ongoing trend. It began when functional departments within organizations began to see themselves, not as separate entities with independent informational needs, but rather as part of an integrated whole which should work together to create an integrated information base. This trend progressed further as organizations realized the benefits of reaching out beyond their

organizational walls, and began exchanging information with their economic partners, such as suppliers and customers. The participants in this project have extended the concept of information sharing “one step further” and have demonstrated the economic benefits of cooperatively developing and exchanging data among organizations which are essentially independent, but which have overlapping informational needs.

The Internet-based Clearinghouse has proven to be a practical tool for fostering the desired cooperation and achieving the desired benefits. It is being used today with the tools currently available and will most certainly evolve into a more powerful instrument. Discussions are underway to transfer the Clearinghouse to a state agency, such as the State Library, for permanent operations.

Appendix A: Related Products

NYS Department of Environmental Conservation, *Development of a NY Geographic Information System Cooperative*, CTG.GIS-001, May 1994.

Kelly, Kristine. *Compelling Reasons for GIS Coordination in NYS*, CTG.GIS-004, December 1995.

Kelly, Kristine. *A Framework for Evaluating Public Sector Geographic Information Systems*, CTG.GIS-005, December 1995.

Kelly, Kristine, Pardo, Theresa, Dawes, Sharon, DiCaterino, Ann, Herard, Winsome. *Sharing the Costs, Sharing the Benefits: The NYS GIS Cooperative Project*. CTG Project Report 95-4, December 1995.

Appendix B: Project Participants

NYS Department of Environmental Conservation

Lawrence Alber, Manager, GIS Unit

Thomas Donovan, Director, Division of Information Systems and Management Planning

NYS Forum for Information Resource Management

Terrence Maxwell, Ph.D., Executive Director

Meghan Kiernan, Graduate Intern

Erie County Water Authority

Paul Becker, Distribution Engineer

Philip Cook, Ph.D., Administrative Director

University at Buffalo, National Center for Geographic Information Analysis (NCGIA)

Hugh Calkins, Ph.D., Director

Center for Technology Government

Anthony Cresswell, Ph.D., Faculty Fellow, School of Education

Sharon Dawes, Ph.D., Director

Ann DiCaterino, Manager, Project Support

Sally Goodall, Administrative Assistant

Winsome Hérard, Assistant Project Coordinator

Fazal Ilahi, Graduate Assistant, Department of Computer Science

Kristine Kelly, Research Associate

Anne Miller, Graduate Assistant, Department of Geography and Planning

Theresa Pardo, Project Coordinator

Division of the Budget

Paul Fisk, Deputy Chief, Technology Management Unit

Michele Lachiusa, Senior Examiner

State Archives and Records Administration

Thomas Ruller, Archivist III

Junlei Zhang, Senior Public Records Management Specialist

Daniel Lorello, Archivist III

Elizabeth Maio, Senior Public Records Management Specialist

Stanley Schwartz, Local Government & Technology Services

Corporate Partners

Applied GIS, Inc.

Aule-Tek, Inc.

Blue Moon Training Systems

Digital Equipment Corporation

Documentation Strategies, Inc.

Full Circle Communications

Harlan Wallach Graphic Arts

Inteligis Corporation

Sun Microsystems Computer

Xyplex

Appendix C: NYS Spatial Data Clearinghouse Technical Information

Environment

The NYS Spatial Data Clearinghouse exists as part of a larger spatial data clearinghouse environment. The National Spatial Data Clearinghouse (NSDC), an initiative of the Federal Geographic Data Committee, is a network of virtual and physical repositories of spatial data available over the Internet. The url for the FGDC home page is <http://fgdc.er.usgs.gov/>

Hardware

The Clearinghouse was implemented within CTG on a DECstation 5000 running Ultrix 4.2, with 1 gigabyte of storage and 128 megabytes of memory.

Contents

Forty-nine spatial data sets are described in the Clearinghouse with metadata which conform to the Federal Standard for Digital Geospatial Metadata.

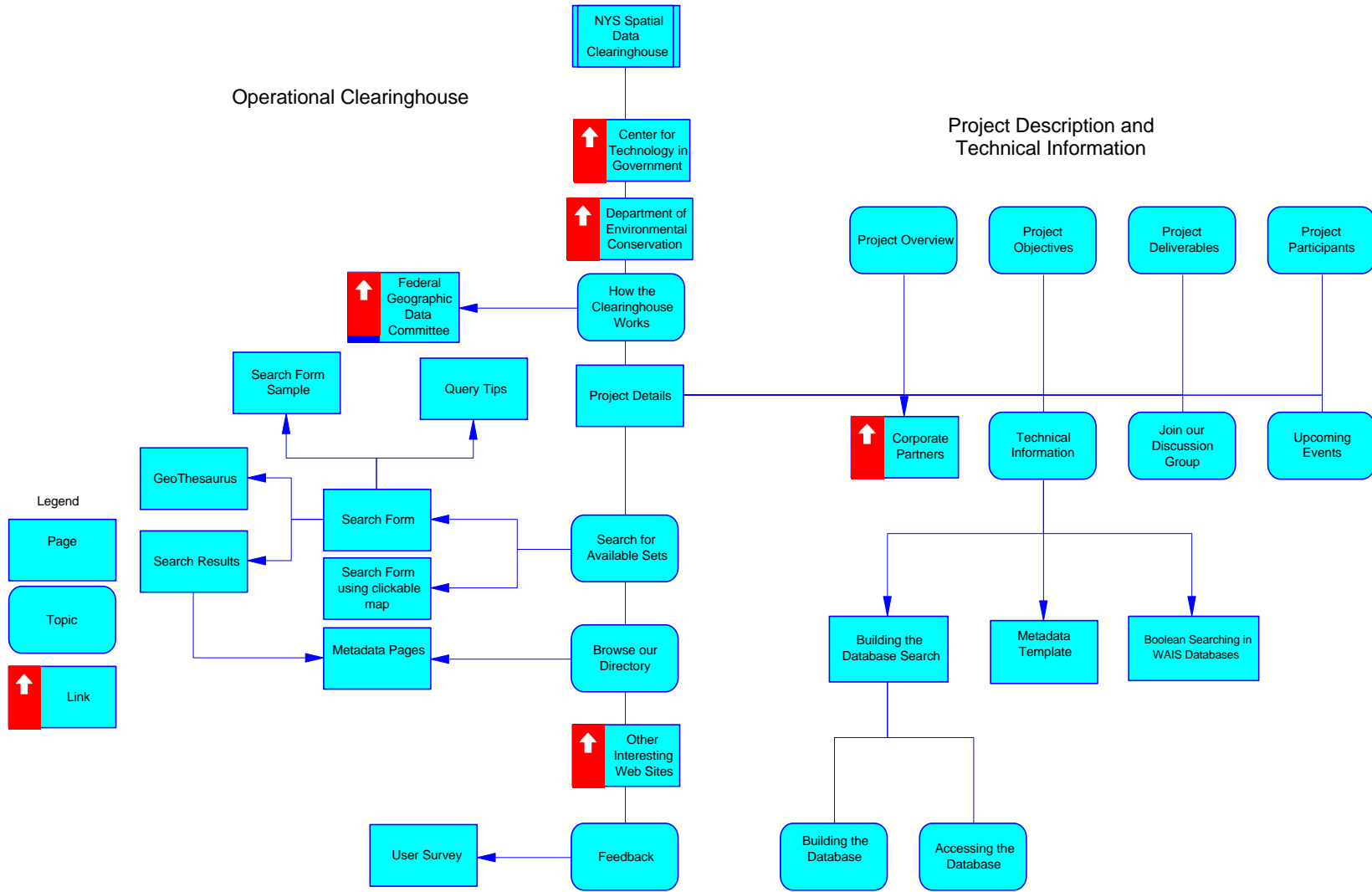
The Federal Metadata Standard was adopted as the standard for the Clearinghouse. Organizations interested in submitting data were provided with a packet containing information about the standard and a template of the standard on disk with embedded html code. The template was provided to support the collection of the metadata. In addition to providing metadata some users provided both the data set and a image of the area represented by the data set.

Structure

The figure on the following page represents the structure of the prototype Clearinghouse.

NYS Spatial Data Clearinghouse

Information Structure



The following pages provide information on the operation and construction of the Clearinghouse. These pages are reproduced from the Home Page of the Clearinghouse and from the Technical Details section.

**Center for Technology in Government
University at Albany / SUNY
1535 Western Avenue
Albany, NY 12203
Phone: 518-442-3892
Fax: 518-442-3886
info@ctg.albany.edu
www.ctg.albany.edu**