STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE

Working Paper No. 8
English only

CONFERENCE OF EUROPEAN STATISTICIANS

UN/ECE Work Session on Methodological Issues Involving the Integration of Statistics and Geography
(Neuchâtel, Switzerland, 10-12 April 2000)

Topic (ii): Spatial database management and (geo-)data warehousing

BUILDING A GEOSPATIAL DATA CLEARINGHOUSE FOR DATA DISCOVERY AND ACCESS

Submitted by U.S. Federal Geographic Data Committee[1]

Invited paper

ABSTRACT

The Geospatial Data Clearinghouse has been operating in the United States since 1996 to provide a discovery service for digital geographic information of all types. Primarily designed to facilitate sharing of data collected and managed by U.S. Federal government activities, the Clearinghouse technology has been widely deployed at all levels (government, private, and academic) in the U.S. and other countries, linking prospective GIS and imagery data users with data providers of all types. This article describes the history, content, and direction of the Geospatial Data Clearinghouse designed and implemented by the U.S. Federal Geographic Data Committee.

I. BACKGROUND

1. In 1994, the U.S. National Spatial Data Infrastructure (NSDI) was officially launched to coordinate the geospatial data collection and management activities between governmental and non-governmental organizations in the United States. There was concern that federal agency digital map data collection was inefficient due to lack of planning and coordination. As the federal budgets were being decreased in the early 1990's and the data collection capacity and requirements of state and local government were increasing, the opportunity for collaborative work between the federal and non-federal sectors was ever more important. One of the primary goals of building a virtual spatial data infrastructure was to improve knowledge about current and future data collection strategies. These strategies do not rely upon outright data sharing, but rather for sharing costs in data collection and maintenance, agreements to exchange data using common standards, and other innovative methods.

2. There is a common misconception that the American NSDI is predicated on the publication of free data. Although it is the practice of the federal government to make all types of information collected at public expense available at the cost of dissemination, other requirements may exist in commercial, local and state government to recover some of the data collection costs. In consideration of this, the U.S. NSDI is a hybrid activity that includes free discovery services, innovative partnerships among data collection organizations, and online, standards-based access to geospatial information through the National Geospatial Data Clearinghouse.

[1] Prepared by Doug Nebert.

II. EXECUTIVE ORDER

3. Through an “Executive Order” of the President to the Executive Branch of the Federal Government, the directive was passed to implement a National Spatial Data Infrastructure. Executive Order 12906 was issued on April 13, 1994 with the following essential definitions (Executive Office of the President, 1994):

(a) "National Spatial Data Infrastructure" ("NSDI") means the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data.

(b) "Geospatial data" means information that identifies the geographic location and characteristics of natural or constructed features and boundaries on the earth. This information may be derived from, among other things, remote sensing, mapping, and surveying technologies. Statistical data may be included in this definition at the discretion of the collecting agency.

(c) The "National Geospatial Data Clearinghouse" means a distributed network of geospatial data producers, managers, and users linked electronically.

4. These words, intentionally general in their scope, have allowed the development of a general purpose solution to spatial data sharing. Unfortunately, the loose definition of geospatial data has allowed some agencies to regard their information as not being geospatial in order to avoid having to comply with the Executive Order. Unlike a Public Law, Congress is not compelled to appropriate specific funds for an Executive Order. There is limited power over agency compliance or enforcement by any authority. Despite these inhibitors, there are a large number of participating organizations who recognize the benefits of participating in Spatial Data Infrastructure activities. These benefits include free advertising, sharing of information, use of low-cost or free enabling software, and cost savings through common data development approaches.

5. The Geospatial Data Clearinghouse is central to the establishment of the NSDI, and was also specifically mentioned in the Executive Order document as an electronic catalog service that hosted descriptions of spatial data sets conforming to the FGDC Content Standard for Digital Geospatial Metadata, also published in 1994. U.S. Federal agencies are required to post their metadata in a searchable Node of the Clearinghouse; GIS professionals in those agencies then should consult with the Clearinghouse holdings prior to the collection or production of new geospatial data. In this way, geospatial data sets can be used for multiple purposes beyond their intended initial scope.

III. NSDI COMPONENTS

6. To better appreciate the role of the Clearinghouse, it is important to understand the interrelation of the components present in the NSDI activity. These elements are not independent activities – they must be viewed as complementary to constructing a viable infrastructure. Figure 1 (see next page) illustrates the essential parts of the NSDI as built to-date.

7. In the figure, each box represents a primary activity being undertaken by members of the FGDC. Lines between the boxes illustrate necessary information linkages between the activities. The Clearinghouse activity provides the primary interaction between users and data. The Clearinghouse services are populated with metadata, or spatial data set descriptions. The metadata also includes linkages to the data or ordering methods that can be accessed by humans or software. Data being described by metadata here is differentiated into two different classes. Framework data consists of well-known themes and data organization of common re-use potential. Other geospatial data which do not conform to well-known specifications are also supported for discovery and access and form the majority of information served within the NSDI. The Clearinghouse supports access to both types of data through their metadata. Adherence to standards developed by the FGDC and other standardization bodies provide a reliable basis for interacting with the components of the NSDI. These parts form the conceptual foundation of the NSDI.

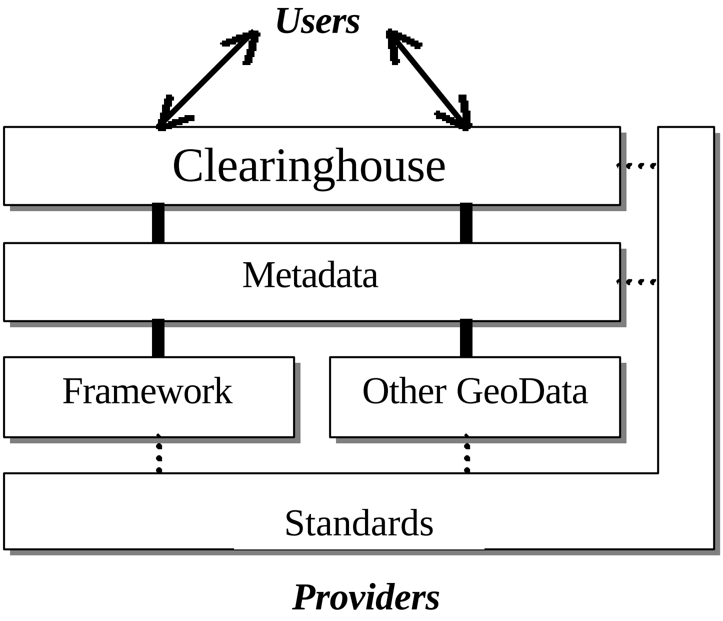

Figure 1. Relationship diagram of the essential activities of the National Spatial Data Infrastructure in the United States

IV. CLEARINGHOUSE CONCEPTS

8. For the NSDI, the Clearinghouse provides a virtual consolidated information space across which searches may be conducted through a single query. A user interested in locating geospatial information fills out a single search form, specifying queries against geographic, temporal, and text fields as well as free-text while targeting one, several, or all registered servers. There are a variety of user interfaces available to the Clearinghouse, but all use the same search approach. This is accomplished through use of a common descriptive vocabulary (metadata), a common search and retrieval protocol, and a registration system for servers of metadata collections.

9. The Clearinghouse assumes distributed ownership and participation. Similar activities on the Internet have taken a fully centralized approach to metadata management by placing all metadata in an index on one server, or in several replicated servers. In an increasingly dynamic data management environment, the synchronization between metadata and the index becomes increasingly difficult. This problem is experienced on a daily basis when conducting searches on Web search engines and getting a “404: File not found” error when a document has been moved or changed. In addition we are seeing a migration toward treating metadata and data as interrelated and even being managed together within a single database. To duplicate this metadata in an external index can be costly and invites problems with synchronization of the data, its metadata, and the externally indexed metadata. Organizations who already manage spatial data and are interested in publishing it are the most capable candidates for publishing and maintaining the metadata. Metadata co-located with data on a server tend to be more current and detailed than metadata published to an external index (harvested and indexed off-site).

10. The Clearinghouse is more than just a catalog of locator records. The Clearinghouse includes data, ordering mechanisms, map graphics for data browsing, and other detailed use information that are stored in the metadata entries. This metadata acts in three roles: 1) documenting the location of the information, 2) documenting the content and structures of the information, and 3) providing the end-user with detailed information on its appropriate use. A traditional catalog, as found in the modern library, provides only locational information. In the era of digital data, the edges between the data and the catalog

can become blurred and permit the management of extended information called metadata that can be exploited by computer software and human eyes for many uses.

V. HOW DOES THE CLEARINGHOUSE OPERATE?

11. A typical session would start with a user interested in finding geographic information for use in their GIS, image processing, or mapping application. A search form is downloaded from one of several NSDI Clearinghouse gateways that lets the user define the geographic area and time period of interest, search against text fields or full-text, and select which servers to search. The user builds a query and passes it to the gateway web server where it is turned into requests to many Clearinghouse servers. Results are returned, eventually as HTML documents to the client as titles of metadata entries that meet the search criteria. The user selects the link associated with a headline and is presented with an HTML (or optionally SGML, XML, or text format) document of the detailed metadata with embedded links they may traverse to download or order the data. The architecture required to make this work is described in the next section and is illustrated in Figure 2.

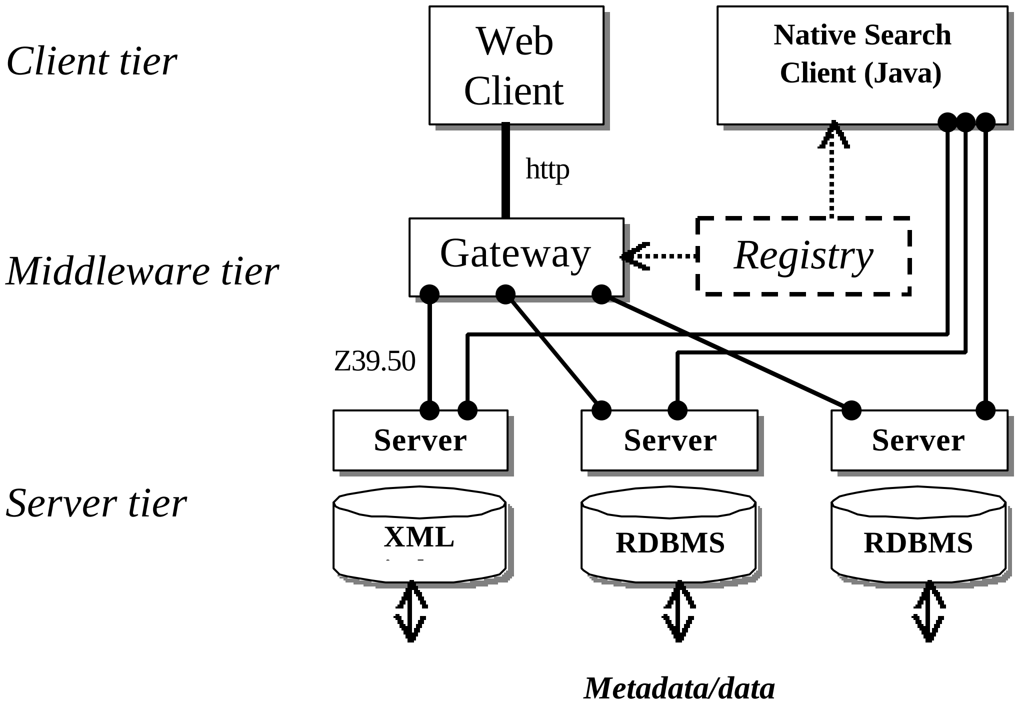

Figure 2. Service architecture of the Geospatial Data

VI. IMPLEMENTATION

12. The Geospatial Data Clearinghouse is implemented using a multi-tier software architecture that includes a Client tier, a middleware or “Gateway” tier, and a server tier, as is illustrated in Figure 2. The client tier is realized by a traditional Web browser or a native search client application. The Web browser uses conventional HyperText Transport Protocol (HTTP) communications, whereas the native search client uses the ISO 23950 (ANSI Z39.50) protocol directly against a set of servers. A commercial Java-based Clearinghouse client, “Meta-Data Browser,” was released in September 1999 by MapInfo to provide desktop access to Clearinghouse servers through a map and tab based search design.

13. The middle tier in the architecture includes a World Wide Web to Z39.50 protocol gateway available commercially from Blue Angel Technologies. Similar functionality may also be achieved through free reference implementation with the 2.07 version of the Isite software suite. Both gateway solutions provide parallel distributed search of multiple Z39.50 metadata servers from a single client Web session. At present, six Gateways have been installed in the U.S. to provide regional points of access. The forms and interfaces installed at each are identical, and each hosts parallel search of all servers. In order to track a large number of Clearinghouse servers, a list of known, compatible servers must also be managed. This list is actually server or collection-level metadata that can itself be searched as a special Clearinghouse catalog. In this way, an intelligent one pass search of eligible servers can be performed instead of requiring the user to select servers from a list, or to have all queries passed to all servers.

14. At the bottom tier of the service architecture are the Z39.50-compatible servers. These servers communicate using a dialect of the Z39.50 protocol named the “GEO” Profile. The GEO Profile is an extended set of the traditional bibliographic fields that can be searched, and includes geospatial coordinates (latitude and longitude) and temporal fields in addition to free-text (e.g. search for the word anywhere in the metadata entry). These Z39.50 servers are implemented on top of XML document databases or relational database systems in which structured metadata are stored for search and presentation.

15. The Z39.50 protocol was selected for use in the Geospatial Data Clearinghouse for several reasons. First, the library catalog service community existed with relevant software and specifications that could be enhanced for geospatial search. By adopting compatible terms, library catalogs can be searched with GEO catalogs. Second, the Z39.50 protocol specifies only client and search behavior and does not specify the native data structures or query language used to manage the metadata behind the server. This abstraction of query allows for a public query on “well known” fields that can be translated at each server into local equivalents. This common search functionality across hundreds of servers is a prerequisite to distributed search. It allows for local database management autonomy yet supports federated search. Third, the protocol is independent of computer platform. Z39.50 search clients and servers exist for many types of UNIX and Windows platforms, and Java libraries are available for additional client and server programming.

16. The separation between local and public metadata search fields has allowed for the Z39.50 search of many different types of metadata collections that support the GEO Profile, even though they may not support the same metadata model. For example, The Australia and New Zealand Land Information Council (ANZLIC) metadata contains different tag names than FGDC metadata. Through special translation tables in the server, search against ANZLIC's "Data Set Name" field was associated with "Title" as attribute number 4 in the registered public fields. As a result, Australian catalog servers can be searched through the FGDC sponsored Clearinghouse but return metadata records of a different structure. The same approach could be applied to other community metadata services, such as those employed by the Directory Interchange Format (DIF) files used in the space and global change disciplines or other metadata standards with similar content. Ideally, metadata formats should be delivered in such a structure that they could be converted or translated for consistent presentation, even if they come from different communities. The Extensible Markup Language (XML) and translator software may enable this in the near future.

17. To encourage widespread participation in the Clearinghouse, metadata collection and catalog service software has been developed under direction of the FGDC and other coordination organizations around the world. Reference implementations of software exist to provide a free or low-cost example of metadata management and Clearinghouse service that can be quickly implemented. The software can also be used as reference by commercial developers to test anticipated functionality and interoperability and to develop value-added products. The CorpsMet-95 metadata software was commissioned by the U.S. Army Corps of Engineers for detailed collection of full FGDC metadata elements and is available free of cost from the Internet for use on Windows platforms. The MetaLite Windows software was written by the U.S. Geological Survey for use in their international program and manages a subset of FGDC metadata

with interfaces and help available in four languages: English, Spanish, Portuguese, and French. The Isite software suite is a reference implementation of the Clearinghouse server that includes an XML/SGML-based indexing system and a Z39.50 server for use on Windows and UNIX platforms. Known software tools are inventoried at the FGDC web site (FGDC, 1999). Although free software is useful to the building of a community, the development of high-performance, high-reliability software is required from the commercial sector – especially software that is integrated with GIS software packages on the server and desktop.

VII. STANDARDS AND INTEROPERABILITY

18. The Clearinghouse has been designed with maximum reliance on existing technologies and standards. Because of this, existing software could be re-utilized or adapted to support geospatial information without requiring special investment in new technologies. The use of a generalized query protocol on Z39.50 will permit a migration from FGDC metadata to future forms being developed through international consensus under ISO Technical Committee 211 and their draft metadata standard 19115. Even though the metadata standard will change, the GEO Profile specifies the meaning of search fields in a way they can be mapped to multiple metadata schemas. Under the GEO Profile we can support search of metadata in Europe (Global Environmental LOcator Service, GELOS), the United States, Canada, Latin America, and Australia in a single search, even though different metadata models are applied.

19. The OpenGIS Consortium (OGC) solicited a catalog services specification in October 1998. A group response was submitted from OGC members Oracle, ESRI, Intergraph, Blue Angel Technologies, FGDC, and the National Imagery and Mapping Agency and others in July 1999. At the August OGC meeting in Southampton, U.K., a common catalog services approach was presented and demonstrated that built upon the essential search and retrieval model of Z39.50. Implementation specifications were submitted for CORBA, OLEDB, and the Web. Distributed parallel search across these different protocols was shown through gateway software. The Web Profile of the OGC Catalog Services Specification includes two implementation paths: one permits the implementation of existing Z39.50 servers (on TCP/IP) and a second encourages the use of XML encoding of queries and responses but over HTTP. These XML Encoding Rules (XER) were demonstrated at Southampton through client and server software developed by the Joint Research Centre of the European Commission and provide an interesting alternative to traditional Z39.50 approaches. Because the server is implemented on HTTP, metadata providers need only install the server and index software as part of their web server and not worry about an additional process or protocol. Firewall issues of using a different TCP port are minimized because all queries could use the web server's communication port. CORBA and OLEDB implementations provide solutions for organizations that are already using those technologies. Gateway technology will be needed to support distributed search across all protocols and/or bridges to translate query and results between them.

VIII. ISSUES FOR THE FUTURE

20. The goal of global spatial data discovery and access will not be achieved without developing common standards and practices in building spatial data infrastructures. The Global Spatial Data Infrastructure (GSDI, 1999) activity is an international effort to raise awareness about the supportive policies and technical standards that promote compatible geospatial data access. Rather than invent new standards, the GSDI participants are identifying best practices and standards being developed in regional and international settings that can be applied at all scales of application. The Technical Working Group of the GSDI is developing a dynamic electronic document that will assist countries and organizations in developing policies and technology that are compatible with national and global infrastructure initiatives. This document is scheduled for initial release at the next GSDI conference in Cape Town, South Africa in March 2000.

21. The Clearinghouse strategy bridges many different disciplines and relies heavily on the existence of a current registry of Clearinghouse servers. Without such a current list of metadata servers, searches beyond ones traditional set of information resources would be limited. Over 150 servers are now registered with the U.S. Clearinghouse -- a number that doubles every year. Searches of the Clearinghouse by software or humans should be able to rely on a current and exhaustive list of available servers, not just in the US or FGDC-affiliated collections, but worldwide. The development of an international registry of metadata services is being developed under GSDI for use in national, regional, and global search applications. This should come online in early 2000.

22. The production of metadata is not a task that has yet been integrated into the mainstream GIS and image processing software suites. In most cases, metadata collection is viewed as an exercise separate in space and time from GIS data management. Metadata and spatial data must be viewed as interrelated aspects of a geographic data set and managed together for the entire lifecycle of the data product. As data are produced, so are metadata; as data are edited, so should metadata; when data are published and exported, so always should the metadata travel as part of the data package. The inclusion of metadata management tools (ingest, edit, export, index, serve) within desktop and professional GIS and image processing products would greatly improve the ability to publish geospatial data using Clearinghouse concepts.

23. Semantic mappings will be needed among keywords used in different communities and languages. Search across hundreds of servers in dozens of languages will be difficult and imprecise unless conventions are established to support multilingual search. Clearinghouse search uses geographic coordinates (latitude and longitude) and dates for spatial and temporal search that would work in any language. The discovery of documents of a certain topic or thematic content, however, will require the use of international thesauri of terms by which all metadata could be classified. The European Environment Agency's GEneral Multilingual Environmental Thesaurus (GEMET) offers over five thousand terms in 13 languages. Unless a global community can select a single thesaurus or classification system, bridging several multi-lingual classification systems will still be difficult. This is an issue requiring consistent international policy that, once resolved, can be addressed by software solutions. In the future, search in any language could yield reliable "hits" on data sets described in other languages. Automatic linguistic translation of documents may still be years away, but their discovery is feasible in today's software when using a single, agreeable multilingual thesaurus.

IX. SUMMARY

24. The Geospatial Data Clearinghouse developed for the FGDC community is in its fifth year of operation. Its primary goal is to support discovery and access of geospatial data resources in the United States and other countries so that duplication of effort in data collection is less likely to occur. Over 150 Clearinghouse servers now exist with data for all continents. The Clearinghouse acts as the primary public access point for all resources managed within the NSDI. The Clearinghouse employs a distributed architecture that permits search of many servers through a single interface with a user interface style familiar to those used in Web search engines. Use of the Clearinghouse does not require special software -- web browsers and special web servers known as Gateways provide access to the distributed collections. Although the Z39.50 protocol has been used in Clearinghouse to preserve generalized query across many servers, updated specifications recently passed by the OpenGIS Consortium extend these capabilities to the Web and other protocols that can support search against FGDC, ISO TC 211, and other metadata models. Many agreements are needed on building national and international infrastructures. Hopefully through efforts like those of the Global Spatial Data Infrastructure, the process will be an easier one leading to greater compatibility across national borders for solving problems that transcend the local environment.

REFERENCES

Executive Office of the President, 1994. "Coordinating Geographic Data Acquisition and Access: The National Spatial Data Infrastructure," Executive Order 12906, April 13, 1994, Federal Register, Volume 59, Number 71, pp. 17671-17674.

Federal Geographic Data Committee, 1999, FGDC Home Page: http://www.fgdc.gov.

Global Spatial Data Infrastructure, 1999, GSDI Home Page: http://www.gsdi.org.