National Spatial Data Clearinghouse: The Real Story

By Peter N. Schweitzer

U.S. Geological Survey
Mail Stop 954
Reston VA 20192
Telephone: 703-648-6533
Fax: 703-648-6252
e-mail: pschweitzer@usgs.gov

INTRODUCTION

Since 1995, Federal and state agencies have put a lot of work into the National Spatial Data Clearinghouse. While that work has been fruitful, we've learned some useful things by looking closely at the character and usage of the clearinghouse and of the metadata it contains. The distributed search system designed for the national clearinghouse, though functional, is little used by the public it ostensibly serves. Are its contents therefore irrelevant? By no means! Usage statistics from a well-monitored clearinghouse site reveal that the public overwhelmingly prefers to use standard web search tools and local site navigation to find information. These observations support a new view of data catalog presentation that relies less on centralized search infrastructure, building more on the consistency inherent in metadata, increased use of controlled index terms, greater innovation in presenting information, and monitoring of actual use.

Geological surveys and other scientific organizations increasingly recognize the value of spatial data and the importance of well documented digital data for users both within and outside their walls. Efforts to communicate spatial data and metadata that cross organizational boundaries thus represent a common interest and a potential avenue for increasing the efficiency and improving the usefulness of the results of their research and monitoring. Once these efforts pass beyond the planning and promotional stages of development to operation, however, it becomes possible and important to evaluate their effectiveness in practice.

This report is intended to assess the effectiveness of the National Geospatial Data Clearinghouse search system from the point of view of an organization that distributes scientific data and metadata of broad appeal to the public.

BACKGROUND

With Executive Order 12906, signed by President Clinton in 1994, the National Spatial Data Infrastructure (NSDI) became a significant goal for Federal agencies and their funded cooperators, to which other organizations were encouraged to contribute. Briefly, the NSDI required data produced by Federal agencies to be documented using the metadata standard developed by the Federal Geographic Data Committee (FGDC), an interagency group charged with coordinating spatial data activities of, and cooperation among, Federal agencies. The metadata produced by the agencies was to be stored in a clearinghouse whose design was not finely specified in the order, and agencies were required to consult the clearinghouse before spending money to gather or purchase spatial data. The executive order also described the creation of the national geospatial data framework and stressed partnerships and standards development as well. This paper focuses on the spatial data clearinghouse.

Charged with the development of a clearinghouse that agencies would consult, yet not given the mandate to centralize spatial data delivery, the FGDC developed an architecture for the clearinghouse in which the metadata from different organizations would be stored and maintained in computer systems distributed widely across the country. Using internet protocols well established in the library computing community, the FGDC would devise search interfaces that would communicate requests from users to these widely-dispersed servers and, upon receiving results from them, pass the search results to the person who requested the information. This architecture is illustrated diagrammatically in figure 1.

The following factors that motivated FGDC to choose this clearinghouse architecture are important for the present discussion:

- People don't know where the servers are
- People don't know how to search each server if each server shows a different interface
- Why should everyone build a different search interface?

These concerns were and remain valid reasons to support, in principle, the distributed clearinghouse

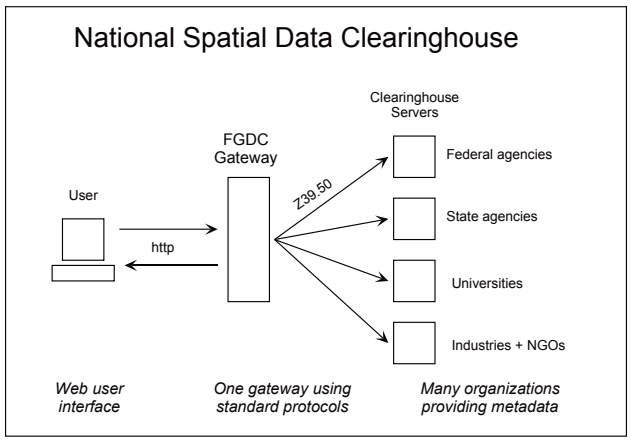

Figure 1. Diagram showing the architecture of the National Spatial Data Clearinghouse. The user wishing to search multiple metadata sources enters a dialog with the gateway machine, which communicates search requests to each of the clearinghouse's metadata servers according to the Z39.50 protocol, which is well-established in the library community. Responses are returned to the gateway and then to the user.

architecture with its centralized search interfaces on "gateway" systems.

CURRENT STATUS

Supporting the clearinghouse are many people engaged in several rather different activities. Writing metadata is perhaps the best-known of these, because it confronts scientific and technical experts with the unfamiliar and sometimes daunting structure of standard metadata. An organization with more than a few data producers will find that it needs to dedicate attention to the task of gathering metadata; arranging them in a collection; and imposing consistency in the expression of common terms such as publication series, keywords, disclaimer statements, format descriptions, and network addresses. The same people typically configure and run one or more clearinghouse servers, computers that understand the Z39.50 internet protocol used for searching distributed collections; and they often administer web servers distributing the same information using hypertext transfer protocol. Farther along the chain are those who design, set up, and run gateway systems, which provide the search interface that people can use to query the clearinghouse.

This report focuses on the concerns of the metadata collection manager and clearinghouse node administrator. In this role the chief concern is that the information to be provided to the public can be found, obtained, and used appropriately. Within USGS I carry out this function for geologic data. I begin this discussion with the frank admission that I find the search interface that FGDC has developed to be cumbersome and confusing. I believe that if alternative interfaces are provided to users they will be employed in preference to the clearinghouse interface.

TESTING THE EFFECTIVENESS OF THE CLEARINGHOUSE

For a person maintaining information to be made available to the public, the most important performance measure is to what extent that information is obtained by the public and the context within which the information is provided.

The USGS Geoscience Data Catalog is a collection of metadata records describing a wide variety of research results produced in the course of geological research conducted by USGS. These records are generally highly detailed and of excellent quality and consistency. At this writing the collection contains 1,589 records, but during the period in which statistics were gathered for this report the collection contained 1,117 records.

The records in this collection are accessible both using the Z39.50 protocol of the NSDI clearinghouse and using the typical hypertext transfer protocol (HTTP) used by most web sites. By examining the number and frequency of metadata records provided using these two protocols, we can learn how effectively they are distributing information to the public. An examination of this sort proceeds by analyzing in detail the log entries produced by the web server and Z39.50 server software.

LOG ANALYSIS

Specialized programs were written for these analyses because general log analysis programs confound{?} several unimportant influences in their results. For example, many requests for documents on the web are the result of automated "robots," also called "spiders," run by general search engines; the spiders read pages and the search engines create indexes of the contents of those pages to assess their relevance to queries that users submit. Similarly, many web sites log requests for files that are not complete documents but are ancillary information such as images. Likewise, web servers write log entries for requests even when the user already has a current copy of the information requested or when the information is known to be on a different server.

The web log analysis program written for this study examines all entries in the HTTP server log and counts only those requests that:

- *ask for bona fide metadata records*
  This is determined by examining the location of the requested file; on this server, metadata records are stored beneath the directory "metadata".
- *represent successful downloads*
  The HTTP server returns a result code "200" for

successful downloads; other codes are indicated for redirection or if the user already has an up-to-date copy of the file.
- *originate from users outside USGS*
  Since I am from USGS, this reduces the likelihood that the statistics are biased by my own download activity.
- *are not from a web spider or robot*
  This is accomplished by watching the user-agent field of the HTTP server logs. Every day a separate file is written containing only this field. By sorting it and eliminating duplicate entries, one can identify by inspection those user agents representing spiders, and store in a separate file a list of text strings that, if present in the user- agent field, indicate a spider.
- *result in complete downloads*
  The log records the number of bytes transferred; this is compared with the actual size of the file requested.
- *are not duplicate transfers*
  If the same host downloaded the same file on the same day, it is a duplicate and only one should count.

Source code for this program is available at:

http://geo-nsdi.er.usgs.gov/stats/yesterday.c
main program analyzes web logs
http://geo-nsdi.er.usgs.gov/stats/spiders.c
subroutines that read database of spiders
http://geo-nsdi.er.usgs.gov/stats/spiders.txt
database of web spiders

The programs are provided here for reference; they would need some customization in order to run on another computer system.

Metadata records on this server are available in several different formats: parseable (indented) text, outline-style HTML, Frequently Asked Question (FAQ)-style HTML, Standard Generalized Markup Language (SGML), and Directory Interchange Format. The output of the program `yesterday` shows the number of metadata records of each available format requested by real users outside USGS during the previous day. As ancillary information, the program also creates a file containing only the HTTP user-agent identifiers for these downloads, and another file containing the HTTP referrer, which is, for each download, the address of the web page containing the link the user clicked in order to download the metadata record. The summary of web downloads is updated daily and is available at http://geo-nsdi.er.usgs.gov/activity.shtml.

The Z39.50 server software used at this site is Isite, which was developed by the Center for Networked Information Discovery and Retrieval at the University of North Carolina and is maintained by the FGDC for use with the NSDI clearinghouse.(see http://clearinghouse4.fgdc.gov/ftp/) As distributed, the Z39.50 server software produces logs that are not sufficiently detailed to answer the question posed here; it records the search and present requests and the number of bytes transferred in the present request, but it does not record the type of record requested. Several record types can be requested: full, brief, and summary. Of these, only full records are significant. The brief and summary records provide only the document titles and are used by the gateway to compose a list of relevant documents which the user may request. To carry out this study, I modified the source code of Isite, recompiled it, and use the modified version to obtain more complete logs in Extensible Markup Language (XML) format. These modifications have been passed on to the maintainer of the server software, Archie Warnock, who has incorporated them into the current release. The modifications to Isite to improve logging are described at http://geo-nsdi.er.usgs.gov/stats/isite.html.

The modified Z39.50 server log is analyzed using a script written in PHP: Hypertext Preprocessor (PHP) (http://www.php.net/). The script creates an HTML document containing several tables:

- *Summary of activity*
  For each day, the number of search and present requests, the number of errors (mostly search requests specifying database fields for which we have no index or which are not present in the metadata), and the number of records presented, with brief and summary records tallied separately from full records.
- *Error messages*
  For the entire period spanned by the log, shows the number of times each error occurred. Few of these represent misconfigured requests; most are searches on fields that we do not have, such as the international standard book number (ISBN).
- *Clients generating present requests*
  For each remote computer that requests data, the number of requests, with the present requests broken down by type of record.
- *Clients generating search requests*
  For each remote computer carrying out a search, shows the number of searches and the number that could not be carried out due to errors.
- *Fields searched*

For the entire period spanned by the log, shows the number of search requests for each field. Some fields requested frequently are not contained in the database, so these requests cannot be filled. This script is run twice daily and the results are available on the web at http://geo-nsdi.er.usgs.gov/zlog.html

In order to keep the log to a manageable size and the analysis to a manageable time, the log file is changed each

month so that the current log contains only activity from the current month; logs from previous months are stored separately.

RESULTS

To evaluate the relative effectiveness of the Z39.50 and HTTP mechanisms, we compare the number of full records downloaded through the Z39.50 server with the number of metadata records downloaded by real users through the HTTP server.

During the 142 day period from 7 January through 28 May 2003, 102,945 (full?) records were downloaded by real users through the HTTP server. In the same time period, 1,180 full records were downloaded through the Z39.50 server. Overall, 87 times more records were downloaded through HTTP than through Z39.50. (These were the statistics reported at the DMT meeting.) From 7 January through the end of August, 2003, there were 142,179 downloads by real users through the HTTP server and 1,589 downloads through the Z39.50 server, giving a ratio of 89:1 in favor of HTTP. More complete statistics through November 2003 are shown in the two tables below.

Table 1. Web access statistics, by month. Columns show the various file formats available; requestors may access a metadata record in any format. Outline and FAQ are styles of HTML. SGML is available but no hypertext links are provided to the files. DIF is the Directory Interchange Format (version 4) of the NASA Global Change Master Directory. Text refers to the simple format in which indentation is used to display hierarchical relationships among metadata elements. Most hypertext links on the web point to one of the HTML formats.

Date	Text	Outline	FAQ	SGML	DIF	Total
Jan-03	1,401	7,493	11,099	13	1,339	21,348
Feb-03	1,478	7,090	11,531	8	1,393	21,532
Mar-03	1,371	5,383	14,187	17	1,102	22,068
Apr-03	1,376	5,924	13,648	18	1,344	22,314
May-03	1,448	5,464	12,339	28	1,362	20,655
Jun-03	1,047	3,667	8,898	22	937	14,575
Jul-03	762	3,027	7,058	23	822	11,696
Aug-03	753	3,030	6,721	14	736	11,257
Sep-03	945	5,076	10,886	27	1,042	17,979
Oct-03	1,209	5,996	13,501	17	1,174	21,902
Nov-03	1,041	5,304	11,706	13	1,119	19,184
Cumulative	12,831	57,454	121,574	200	1,2370	204,510

Table 2. Cumulative statistics for Z39.50 access through November 2003. When a brief record is requested, only the document's title is provided; only requests for full records actually represent interest by a user. Errors are not necessarily improperly phrased requests, but include searches on fields that are not present in the database or which are not indexed.

Date	Requests		Error	Records presented	
	Search	Present		Brief	Full
Jan-03	13,634	705	6,599	6,340	285
Feb-03	14,819	917	6,463	6,149	311
Mar-03	17,576	1,381	7,323	5,178	175
Apr-03	17,292	1,042	7,206	4,842	219
May-03	15,163	826	6,068	4,319	202
Jun-03	14,218	426	6,456	3,082	118
Jul-03	12,318	771	4,586	2,925	108
Aug-03	10,969	395	4,210	3,275	170
Sep-03	13,293	675	5,890	6,410	160
Oct-03	14,109	430	5,271	4,433	83
Nov-03	12,468	413	5,662	3,539	146
Cumulative	155,859	7,981	65,734	50,492	1,977

The HTTP referer statistics for real users provide additional important information: The most frequent referer by a wide margin is the commercial search service "Google", but the second most frequent referer is a browse interface that is local to this data server: http://geo-nsdi.er.usgs.gov/cgi-bin/place

This browse interface allows people to locate metadata records by choosing from a limited set of commonly-known place names that are arranged hierarchically (continents, countries, states or provinces, and counties). This interface could be built only by ensuring that the metadata records used a rigidly consistent set of place keywords. Note that the FGDC metadata standard allows a record to be categorized using terms from many different controlled vocabulary as well as using non-controlled terms; the recommendation that consistent place keywords be used does not prevent other vocabularies—even other geographic terms—to be used as well, but the vocabulary that is used should be identified in the Place_Keyword_Thesaurus field.

DISCUSSION

In the development of the clearinghouse architecture, much attention was paid to the need for users to carry out a single search on numerous distributed servers, and to restrict the search to specific fields of interest. These are reasonable concerns that general Web-search engines cannot be expected to address. Notwithstanding these concerns, however, real people have chosen to use the web in preference to the clearinghouse by a wide margin. This finding implies that clearinghouse administrators who wish to maximize the distribution of their information to real users will:

- make their metadata available through the web,
- allow them to be indexed by common search engines, and
- build local browse and search interfaces.

From this analysis it is reasonable to ask whether continuation of Z39.50 service is cost-effective. The answer depends on several factors, the most important of which will vary from site to site; that is, the cost of administering the Z39.50 server software. In my experience this is not difficult, so I would not recommend that people who are already running the software discontinue it. But from the perspective of maximizing effectiveness, it is clear that the Z39.50 service is not contributing significantly to meeting the needs of the user community.

It should be noted that the large number of metadata records downloaded through the web indicates that this information is desired by users. If the experiment were simply looking at the number of downloads by Z39.50 on a single server, one might infer from a low number of downloads that people simply don't want this type of information. But since the same information is here available by a different method, that conclusion cannot be sustained. People want these metadata records, and they get them through typical web interactions, not through the clearinghouse.

The Z39.50 server is receiving a large number of search requests, yet is receiving few requests for full records. An examination of the search terms gives us some insight into this problem. Many searches appear to be requests for general topics, such as books and both classical and popular music. Indeed the most commonly searched field is the ISBN, or international standard book number; none of our data sets have this identifier. Many of these requests originate in university libraries, judging from the hostnames from which the searches originate. I believe that commercial software commonly sold to libraries is configured to query all available Z39.50 servers with all searches. It is therefore important not to regard the number or frequency of search requests as a measure of the effectiveness of the clearinghouse.

The clearinghouse search works, but people don't use it. In contrast, Web search wasn't expected to work effectively, but apparently it does. While it's tempting to blame the unpopularity of the clearinghouse interface on lack of publicity or the complexity of the search form, the explanation may be simply that people prefer the familiar to the unfamiliar, and that they will use such a system if it works well enough, even though a more complex, less familiar system would be arguably more comprehensive.