Douglas D. Nebert
GIS Coordinator
U.S. Geological Survey, Water Resources Division
445 National Center
Reston, VA 22092
Telephone: (703) 648-5691

USE OF WIDE AREA INFORMATION SERVERS (WAIS) IN THE DISSEMINATION OF EARTH-SCIENCE DATA

The U.S. Geological Survey (USGS) is prototyping the use of wide area information servers (WAIS) to provide earth-science data users with a consistent interface to USGS data holdings. The WAIS system is implemented using public domain server and client software that runs on platforms that include X-Windows-based UNIX[1] workstations, IBM-compatible personal computers, and Apple Macintosh computers. Clients may query information servers across the Internet using a National Information Standards Organization standard (Z39.50) that permits client-to-server communication without requiring user names and passwords. The result of a query may be a portion of a document, a listing of available sources, a computer program, a data set, or virtually any other media that may be transferred digitally. The USGS is implementing a new version of the Earth-Science Data Directory through WAIS technology that will provide access to listings of USGS data sets including the Global Land Information System, the NASA Master Directory, and the spatial data holdings of other federal agencies as they implement WAIS systems.

INTRODUCTION

The U.S. Geological Survey (USGS) is lead federal agency responsible for the management and coordination of non-digital and digital spatial data. The Federal Interagency Digital Cartography Coordinating Committee (FICCDC) was replaced in late 1990 by the Federal Geographic Data Committee (FGDC) with membership from all federal agencies with mapping responsibilities. Committees exist at the Department of Interior and Bureau (e.g. USGS) levels to assure dissemination of information within the organization.

The USGS Geographic Data Committee has tasked a working group to revise the Earth Science Data Directory (ESDD) to include expanded references and to develop automated links between in-house data inventory systems and the ESDD to assure its currency. References from the National

1. Use of trade names does not constitute endorsement by the U.S. Geological Survey

Digital Cartographic Data Base and the Aerial Photography Survey Reference System of the National Mapping Division, the prototype Distributed Spatial Data Library of the Water Resources Division, and a listing of identified digital spatial databases from the Geologic Division were included in this pilot effort.

As the information resources of the three operating Divisions are scattered across many computer systems and are managed by several different types of software, a spatial data information management system was required that could obtain essential spatial meta-data from each of the systems and make the meta-data, snapshot "views" of the data set, or even the data set itself available across the wide-area network.

WIDE AREA INFORMATION SERVERS (WAIS) CONCEPTS

The Information Systems Division of the USGS identified a copyrighted software package that is available in the public domain called a Wide Area Information Server (WAIS). Pronounced "ways," this software system was co-developed by Apple Computer and Thinking Machines Corporation as a mechanism to make very large amounts of loosely structured information available for query across a wide-area network (Stein, 1991). Traditional information retrieval systems such as on-line bibliographic reference tools or structured data bases require use of a specific query syntax or fixed "keyword" fields. WAIS software allows more flexible storage of documents without the requirement for fixed fields by creating an index of all words and phrases in the document and making them available for free-form query.

Information servers using the WAIS software are being used to provide access to health statistics databases at the National Institutes of Health, stock market information managed by Dow Jones News Service and Peat Marwick, and a variety of databases managed by universities. WAIS software allows information to be retrieved free-of-charge but can support and manage data acquisition fees directly. Although limited in its application at present, the WAIS technology shows great promise for making spatial data holdings and computer application programs available to a large user community across an agency's internal ethernet network or, more appropriately, across a wide-area network such as the Internet.

A wide-area information server is a server process that communicates with other servers and clients using ethernet or TCP/IP communications. The Information Retrieval Service Definition and Protocol Specification (Z39.50) specifies the client-server dialog (NISO, 1988). Because communication occurs directly between clients and servers, users do not need login privileges on the remote machines. This does not pose a security problem as the client to server instructions are limited to requests for data that the server is programmed to support. The client-server model is illustrated in Figure 1.

The NISO Z39.50 standard is compatible with the MARC standard, Z39.2 developed for library catalogs by the Library of Congress. Both of these standards have corresponding International Standards Organization (ISO) standards and are recognized by the National Institute for Standards and Technology.

WAIS differs from other information access programs such as *archie*, supported by McGill

University in several respects. WAIS client-to-server communication relies on TCP/IP rather than higher level ethernet communication. WAIS does not require the creation of an anonymous ftp area and information sources may be read from any directory on the server system using a very specific set of commands. Full text documents can be indexed for query by WAIS as opposed to file and path names and synoptic descriptions supported by some systems. In addition, WAIS client and server software work on a large number of computers and supply graphical user interfaces appropriate to the computer being used. Whereas some programs are geared toward sharing software the WAIS system is designed to share collections of information, viewed as "documents" in a variety of formats.

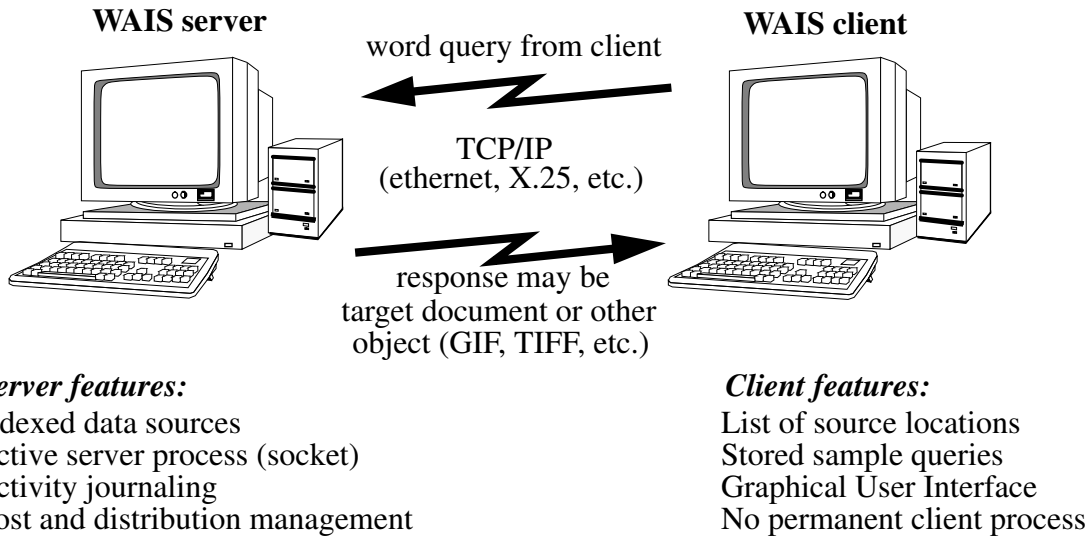

Figure 1. Conceptual diagram of a wide-area information server (WAIS) client-server relationship

IMPLEMENTING A WAIS SERVER

The WAIS software has been written as a set of core routines that build and access server information and a graphical user interface that provides the end-user with access to the candidate information. The core software is written in the c programming language and is highly portable. The user interface modules have been written for the IBM PC under DOS and Microsoft Windows, the Apple Macintosh, X-Windows with and without Motif, and a character-based VT100-style terminal mode. This proliferation of platform support allows the end-user to use his individual computer as the front-end (or client) to remote servers, but also allows the end-user to develop and make available his/her information resources to others on the network. The only requirement for access to a wide area information server is that the computers be able to support TCP/IP -- a capability already prominent in the UNIX workstation world and increasingly available to personal computers, minicomputers, and mainframes.

The first step in developing information to "serve" is to review the intent and content of the data

one wants to make available. The WAIS software allows one to be able to have target information "stores" that are available for query across the wide area network. For traditional information retrieval services this could be thought of as the bibliographic reference, a synopsis, or an abstract, or it could be the whole document. With respect to spatial data, this constitutes the collection of extended meta-data that adequately describe the spatial data set of interest.

Developing a common meta-data exchange format

Because each Division of the USGS manages a slightly different set of meta-data about their data bases, a neutral but robust data format was needed to be the common language of meta-data exchange. After a review of available meta-data collection schemes, the Information Systems Division selected the Directory Interchange Format (DIF) used by the NASA Master Data Directory and the global change community. The DIF format easily accommodates the meta-data produced by each Division despite its primary orientation towards terrestrial and extra-terrestrial image and sensor data. Although the WAIS software does not require that information be managed in specific fields (except for the ability to define a "title" element), it was felt that the presentation of information from the USGS Divisions should use the DIF format to provide a common "face" to USGS spatial data to internal and external users of the data. The *irbuild* function of WAIS was modified by the Information Systems Division to accommodate the NASA DIF format.

An Arc Macro Language program was written to convert meta-data managed by the spatial data documentation program used in ARC/INFO by the Water Resources Division (DOCUMENT), into the NASA DIF format. This program encoded basic information such as the distribution site contacts, distribution media supported, data set description, bounding geographic coordinates, format, and an extended text narrative of how the data were produced into blocks of text that are expected by the DIF interpreter.

The NASA DIF format was designed to store a wide range of data and permits one to tag each holding with one or more discipline values and one or more parameter values. A discipline entry is a hierarchical designation of an area of research through which the data were collected or in which the data might be used. A parameter entry is a hierarchical designation of the object or entity being monitored or, in this case, mapped. The official hierarchies facilitate standardized retrieval of discipline and parameter values at different levels of abstraction, allowing the user to constrain the query to all available data in a discipline or to focus on a class of parameters (e.g. earth science) or a specific parameter (e.g. surficial geology). The parameter hierarchy was extended to accommodate specific types of thematic map data that were not included in the NASA Master Directory. Parameter definitions relevant to terrestrial mapping are given in Appendix 1 with USGS additions to the hierarchy denoted with an asterisk. Note that some types of data, such as elevation data, can be grouped in different ways depending on the way the data are conceptually organized by the researcher. These classifications are not meant to be exclusive but are complementary, allowing the user to discover parameters in alternate hierarchies.

Creating information sources and servers

To create a WAIS "source," as collections of information are called, the DIF files were collected into a single directory and were indexed using the *waisindex* command. This identifies unique

words and phrases and builds an index that clients access through software across the wide area network. An index must be rebuilt each time an entry or update is made to a source collection.

Once an index is built, a server process is created on the host computer that allows local and remote access to the information source. Each source collection must have an active server to make it available to client software. The server is identified by the preferred name of the source, an Internet address, the directory in which the index is stored, a TCP/IP socket number (defined in the /etc/services file), and pricing and access information, as shown in Figure 2. Once the server is active, anyone with wide area network access can query the information holdings.

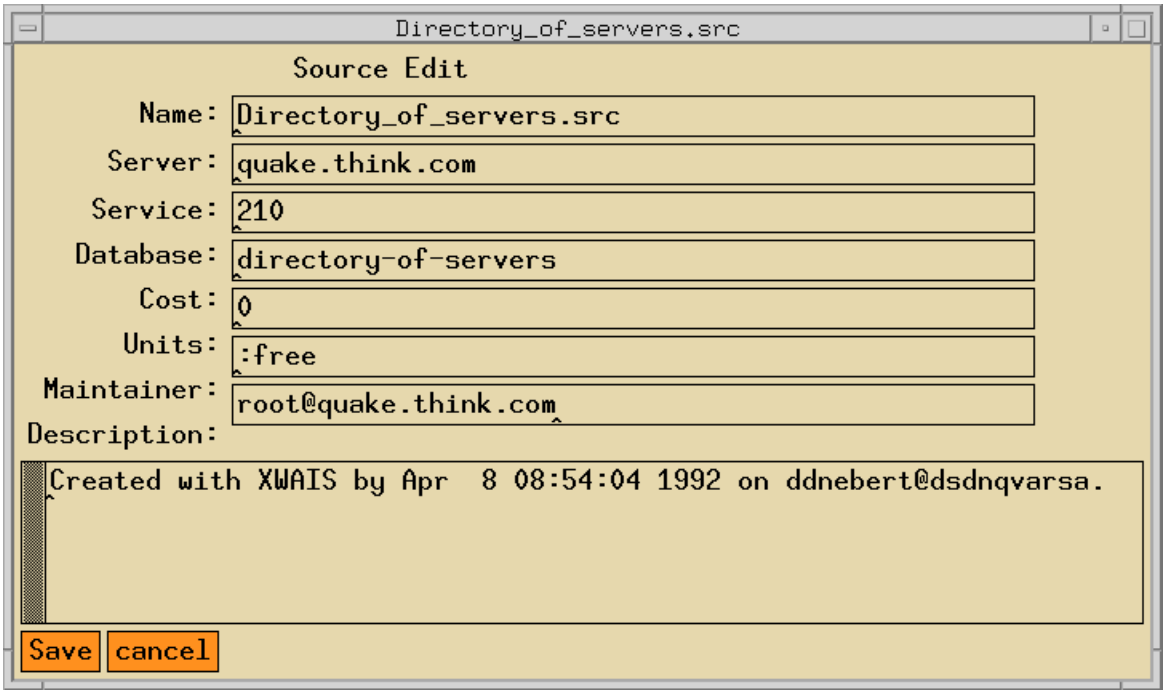

Figure 2. A server entry for the directory of servers maintained by Thinking Machines Corporation.

Thinking Machines maintains a source collection of all known available sources held by a large number of institutions and companies including the University of California, Santa Barbara, University of North Carolina, Southern Illinois University, National Institutes of Health, Columbia Law Library, Apple, Sun, Next, Digital, and the Dow Jones News Service. A user can build a personal list of sources by first posing questions of the Directory of Servers maintained by Thinking Machines. Most of these services are free to the public beyond any usage charges on private networks.

CREATING AND ACTIVATING A WAIS CLIENT

The WAIS software provides a client process that is typically accessed through a graphical user interface. As described earlier, window interfaces have been written for the IBM PC under Microsoft Windows, the Apple Macintosh under MacOS, and X-Windows with a Motif option for UNIX platforms. There is also an interface available through GNU EMACS and a non-graphic

version for VT100-style terminals.

After setting some directory variables (and in X-Windows, merging X resources) the client process can be invoked with the *xwais* command. The initial window that is presented to the user is shown in Figure 3. The window is divided into three optional panels. The first panel displays a scrolling list of saved questions that one might want to re-use. On starting the use of WAIS, this list is empty. The second window is a scrolling list of available source references that may point to local or remote source collections of information. The third panel of the window contains a help and quit button and a message window describing the status of the process.

The first task for the user interactive with WAIS is to create a reference to one or more sources against which queries could be posed. By pressing the *new* button on the source panel an empty form is presented in a new window in the format of figure 2. One may either fill in information about remote sources of information, such as the Directory of Servers (as highlighted), or known WAIS server sources that the user has created or has access to. The reference list of sources may be maintained by each user to reflect their interests. This list is simply a pointer to where the information resource is actually located. The sources may be located on the local computer but are more often accessed from remote systems. Once sources have been entered and saved the client is now ready to process queries.

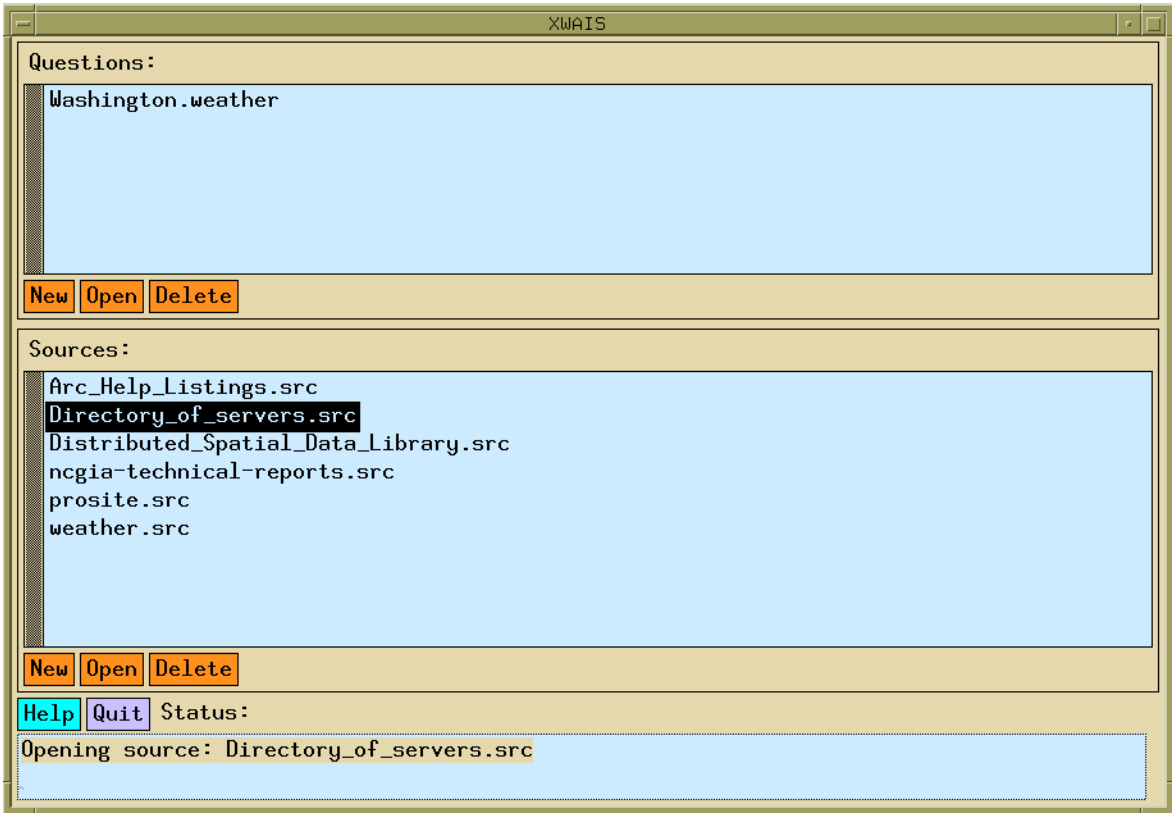

Figure 3. The initial screen of the *xwais* interface showing lists of stored questions and known sources of information.

POSING QUESTIONS OF WAIS

By pressing the *new* button on the question portion of the screen a new window is presented which contains information relevant to a specific question. This question window is shown in figure 4, and is divided into four main fields. The first field, displayed across the top of the window is the query field. It is into this field that words and literal strings are typed for matching against indexed information sources on the wide-area network. The second field shows a scrolling list of target sources to be queried. The third field allows one to limit the query by using selected documents as examples. The fourth field displays the candidate document titles that were found by the search.

The example window shows a search for "streams" out of a list of sources called "Distributed_Spatial_Data_Library.src" which is at present a listing of available digital spatial data at the headquarters node of the Water Resources Division.

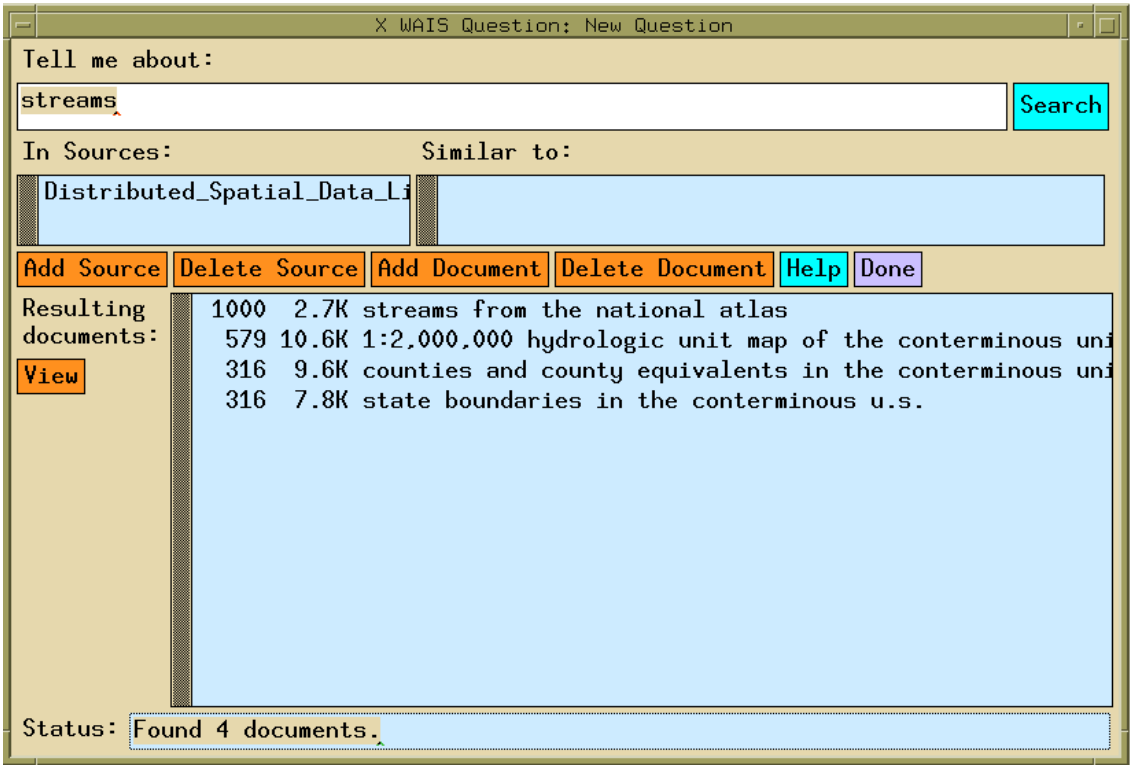

Figure 4. WAIS question window.

After a word or phrase is entered in the question field and one or more sources are added by pressing the *Add Source* button, the search is initiated by pressing the *Search* button in the upper right of the window. The client communicates with each of the necessary WAIS servers and retrieves references to each candidate entry. Rather than automatically retrieving the entire document, the WAIS software retrieves only the reference and provides relevance feedback, calculating a confidence value with which each entry meets your search criteria. This value is listed as the first field in the resulting documents list. The relevance is based on a hierarchy of properties

that can be modified in the search engine. As it is currently set, the occurrence of the target word in the title gets the highest score, and then the frequency with which it is used in the indexed document is evaluated. A value of 1000 is the best match, with lesser values assigned to less "qualified" entries.

Clicking on the entry of interest will cause the WAIS client to fetch the target entry from the server. This target document may be an abstract, an entire text document, a postscript file, or even an entry. It is possible to have more than one of these forms of data made available as the result of a successful query. The candidate entry is then placed in a window for viewing on the screen and may be saved to a local file or deleted from the client system. In the prototype system this retrieves an extensive file of meta data for each selected spatial data theme.

Another feature of the WAIS client is to re-submit a search for candidate entries that are similar to a selected entry. This is particularly helpful if a large number of entries are received from a query. To limit the search, a highlighted document may be added to the search strategy by pressing the *Add Document* button before re-issuing the search.

WAIS IN EARTH SCIENCE INFORMATION ACCESS

The WAIS client-server relationship shows great promise in providing a graphical interface to spatial data holdings. In the current Earth Science Data Directory (ESDD) prototype, references to data have been easily entered into a WAIS server. Interfaces are being written to access the National Digital Cartographic Data Base (NDCDB) so that current holding references are immediately available. The prototype server is available for general public access but is not programmed to automatically retrieve or distribute available data sets.

Interfaces are also being written between WAIS and the NASA Master Data Directory and the Global Land Information System (GLIS). Both these systems manage and provide for retrieval of earth monitoring data such as imagery from around the world. WAIS will direct prospective users to these two systems when candidate data sets are identified there.

The National Center for Geographic Information and Analysis at the University of California at Santa Barbara has created a WAIS server of its on-line technical reports. The full text of the reports is indexed so that searching for very specific terms within their publications is facilitated, returning not just a reference to the report but the entire report in digital form.

The Southern Illinois University at Carbondale acquires and processes hourly weather maps and images received from the National Weather Service. These data are processed and placed in Graphics Interchange Format (GIF) files for immediate incorporation into a WAIS server of weather data. In this way, current surface analysis maps, visible and infrared satellite imagery can be made immediately available to hydrologic researchers and converted for use in GIS software.

The principal benefits of using WAIS software include 1) no requirement to use or be familiar with a given GIS package (except for exchanging meta-data), 2) indexed data can be from a variety of sources and in a variety of formats including graphical images, binary and text files, and 3) user does not need to know the organizational logistics of data access -- a consistent interface to earth

science data is presented.

SUMMARY

Wide area information servers are being tested by the U.S. Geological Survey and other organizations for the dissemination of earth science information. Spatial data holdings, references, and technical documents are being indexed and placed into WAIS servers for public access through windowing interfaces on a variety of computers across the Internet wide area network. The WAIS concept, supported by a library communication protocol, Z39.50 allows clients to access information from local and remote holdings without requiring remote login privileges and passwords. The results of a query can be selected or rejected by the user, reducing overall network loads by only transferring documents of interest rather than all candidate documents. As the result of a query, a server may return text files, synopses, or graphical images to the client for review and further use. This has the potential of making the assessment of spatial data holdings more tangible than a text-only reference retrieval from traditional automated reference systems.

REFERENCES

National Information Standards Organization, 1988. Information retrieval service definition and protocol specification for library applications, Z39.50-1988. Bethesda, Maryland.

Stein, R.M. “Browsing through terabytes.” BYTE magazine, May 1991, pages 157-164.

ADDITIONAL SOURCES

Source and executable WAIS software may be obtained in the following way:

% **ftp think.com**
user: **anonymous**
password: **anything**
> **cd wais**
> **ls**

For UNIX WAIS:
> **type binary**
> **get wais-8-a11.tar.Z**
> **quit**

For Macintosh WAIS:
> **get WAIStation-0-62.sit.hqx**
> **quit**

-- use Stuffit to “Decode Binhex file” then UnStuff the archive

For information about other WAIS, send electronic mail to “wais-interest@think.com”.

APPENDIX 1. EXTENDED NASA DIF PARAMETER HIERARCHY

Climate	
Climate Zone*	Atmospheric Dynamics>Climate Zones*
Temperature	Atmospheric Dynamics>Atmospheric Temperature
Evaporation	Atmospheric Dynamics>Evaporation
Evapotranspiration	Atmospheric Dynamics>Evapotranspiration
Humidity	Atmospheric Dynamics>Humidity
Precipitation	Atmospheric Dynamics>Precipitation
Pressure	Atmospheric Dynamics>Pressure
Sampling Locations*	Atmospheric Dynamics>Sampling Locations*
Solar Radiation	Atmospheric Dynamics>Solar Radiation
Winds	Atmospheric Dynamics>Winds
Hydrology	
Contamination	Hydrologic Parameters>Contamination
Deposition	Hydrologic Parameters>Deposition
Erosion	Hydrologic Parameters>Erosion
Evaporation	Hydrologic Parameters>Evaporation
Glaciers	Hydrologic Parameters>Glaciers
Ground Water	Hydrologic Parameters>Ground Water
Infiltration	Hydrologic Parameters>Infiltration
Model Definition*	Hydrologic Parameters>Model Definition*
Oxygen Demand	Hydrologic Parameters>Oxygen Demand
Precipitation	Hydrologic Parameters>Precipitation
Rivers	Hydrologic Parameters>Rivers
Runoff	Hydrologic Parameters>Runoff
Sampling Locations*	Hydrologic Parameters>Sampling Locations*
Sedimentation	Hydrologic Parameters>Sedimentation
Solids	Hydrologic Parameters>Solids
Surface Water	Hydrologic Parameters>Surface Water
Temperature	Hydrologic Parameters>Temperature
Turbidity	Hydrologic Parameters>Turbidity
Water Vapor	Hydrologic Parameters>Water Vapor
Wetlands	Hydrologic Parameters>Wetlands
Geology	
Age Determinations	Geological Parameters>Age Determinations
Aquifer*	Geological Parameters>Aquifer*
Coal	Geological Parameters>Coal
Economic Minerals	Geological Parameters>Economic Minerals
Geochemical Analysis	Geological Parameters>Geochemical Analysis
Igneous and Metamorphic Rocks	Geological Parameters>Igneous and Metamorphic Rocks
Lithology	Geological Parameters>Lithology
Paleontology	Geological Parameters>Paleontology

Petroleum	Geological Parameters>Petroleum
Petrology	Geological Parameters>Petrology
Sampling Locations*	Geological Parameters>Sampling Locations*
Sedimentary Rocks	Geological Parameters>Sedimentary Rocks
Soils	Geological Parameters>Soils
Stratigraphy	Geological Parameters>Stratigraphy
Surficial Geology*	Geological Parameters>Surficial Geology*
Geodynamics	
Earthquakes	Geodynamic Features>Earthquakes
Erosion	Geodynamic Features>Erosion
Geodesy	Geodynamic Features>Geodesy
Geothermal	Geodynamic Features>Geothermal
Gravity Fields	Geodynamic Features>Gravity Fields
Magnetic Fields	Geodynamic Features>Magnetic Fields
Polar Motion	Geodynamic Features>Polar Motion
Sampling Locations*	Geodynamic Features>Sampling Locations*
Seismic	Geodynamic Features>Seismic
Structures	Geodynamic Features>Structures
Terrain Elevation	Geodynamic Features>Terrain Elevation
Volcanoes	Geodynamic Features>Volcanoes
Geography	
Albedo	Geography and Land Cover>Albedo
Drainage Areas*	Geography and Land Cover>Drainage Areas*
Ecological Zones*	Geography and Land Cover>Ecological Zones*
Elevation	Geography and Land Cover>Elevation
Elevation Array (DEM)*	Geography and Land Cover>Elevation>Grid (DEM)*
Elevation Contours - Hypsography	Geography and Land Cover>Elevation>Contours*
Elevation Points	Geography and Land Cover>Elevation>Point Locations*
Fires	Geography and Land Cover>Fires
Land Use - Land Cover*	Geography and Land Cover>Land Use - Land Cover*
Glaciers	Geography and Land Cover>Glaciers
Ice	Geography and Land Cover>Ice
Lakes	Geography and Land Cover>Lakes
Landforms	Geography and Land Cover>Landforms
Map Quadrangle*	Geography and Land Cover>Map Quadrangle*
Non-Vegetative Surface Features*	Geography and Land Cover>Non-Vegetative Surface Features*
Rivers	Geography and Land Cover>Rivers
Sampling Locations*	Geography and Land Cover>Sampling Locations*
Snow	Geography and Land Cover>Snow
Soils	Geography and Land Cover>Soils

Surface Vegetation	Geography and Land Cover>Surface Vegetation
Surface Water	Geography and Land Cover>Surface Water
Topographic Data	Geography and Land Cover>Topographic Data
Wetlands	Geography and Land Cover>Wetlands
Anthropogenic*	
Manmade Structures*	Anthropogenic>Manmade Structures*
Named Places*	Anthropogenic>Named Places*
Political and Administrative Boundaries*	Anthropogenic>Boundaries*
Census Areas*	Anthropogenic>Boundaries>Census*
Hydrologic Units*	Anthropogenic>Boundaries>Hydrologic Units*
Political Divisions*	Anthropogenic>Boundaries>Political*
Study Area*	Anthropogenic>Boundaries>Study Area*
Populated Places*	Anthropogenic>Populated Places*
Population*	Anthropogenic>Population*
Public Land Survey*	Anthropogenic>Public Land Survey*
Survey Control and Markers*	Anthropogenic>Survey Control and Markers*
Transportation*	Anthropogenic>Transportation*
Roads*	Anthropogenic>Transportation>Roads*
Railroads*	Anthropogenic>Transportation>Railroads*
Trails*	Anthropogenic>Transportation>Trails*
Transmission Corridors*	Anthropogenic>Transportation>Transmission Corridors*