# Locating and Accessing Data and Information on the Internet: Methods and Organizational Impacts

*by Christopher Davis*[1]
*CIESIN*

*"But it's really an information ocean, not a highway. If you think of it as an ocean, then you have to consider the kind of tools that are used, who builds the boats, who designs them, and whether youÆre surfing or diving. If you have a message in the bottle, how do you get the bottle to the people who need it?"* **Peter Gabriel, New York Times, July 13, 1994**

The Internet provides a global infrastructure for data and information publishing that has the potential to revolutionize how data and information are accessed and used. While a variety of methods exist for taking advantage of the capabilities of the Internet for this purpose, two problems with data and information access have been exacerbated by an explosion of tools and resource servers. First, different tools and approaches each have individual advantages, leading to the use of different methods at different locations. Second, resources are distributed across many locations and may often be difficult to locate. CIESIN's information system approach is designed to solve both of these problems by providing a method for locating and integrating heterogeneous, distributed data and information resource servers. The implement of technologies such as those utilized by CIESIN could potentially transform the organization of how data and information are provided and provide users and providers alike with new services and capabilities.

The technical core of the Internet is a set of protocols and standards for computer to computer communication. Initially, Internet technology supported three primary high level services: access to remote systems (telnet), exchange of files (ftp), and electronic mail (smtp). Over time more sophisticated standards and protocols have evolved that offer other services as well. This technology allows the creation of client-servers systems capable of greatly enhancing the dissemination of data and information resources. Data and information providers have a variety of types of resource servers to select from, each with its own functionality (see Table 1).

| **Server** | **Functionality** |
|---|---|
| World Wide Web (WWW)<br>Hypermedia Documentbase | Browsing and retrieval of formatted text, images, sounds, movies combined in hyperlinked documents distributed across servers |
| Gopher Documentbase | Browsing and retrieval of text and/or graphics files in hierarchical lists distributed across servers |
| WAIS Index | Full text searching and retrieval of textual documents and/or graphical files |
| Database | Querying of and access to data structured by fields and records |
| Application | Data processing and analysis of data sets utilizing the capabilities of applications such as ARC/INFO and SAS |
| FTP Archive | Retrieval of text and/or binary files from a hierarchical list |
| Newsgroups/Mailing List | Multi-party, free-form, asynchronous textual discussion |

**Table 1. Internet Server Functionality**

From the user perspective, each of these servers can be accessed

at least by a client application that matches the server.  Some clients, though, provide access to multiple server types (see Table 2).

| | WWW | Gopher | WAIS | Database &[2] Application | FTP Archive | Newsgroup[3] |
|---|---|---|---|---|---|---|
| WWW Browser | **X** | **X** | **X** | **X**[4] | **X** | **X** |
| Gopher Client | | **X** | **X** | | **X** | |
| WAIS Client | | | **X** | | | |
| Database Front End | | | | **X**[5] | | |
| FTP Client | | | | | **X** | |
| News Reader | | | | | | **X** |

**Table 2.  Internet Client Functionality**

Because of the wide range of functionality offered by WWW browsers and servers, these systems are more widely used than any other approach for both access and distribution.

While WWW browsers offer access to a range of servers, the WWW architecture severely limits design flexibility in information systems.  Current WWW standards such as HTML (hypertext mark-up language) and HTTP (hypertext transfer protocol) limit user interface design to what can be accomplished using a basic form interface.  This precludes the use of features such as menu bars and dialog boxes and other dynamic windows. This problem is compounded by the fact that HTTP is a stateless protocol.  The WWW browser requests a document; the document is provided by the server, and the connection closes.  The server has no way of knowing or tracking what document the user requested once the request is filled, thus global settings and variables are difficult to maintain from one document to another. Also, the WWW browser is a display tool only.  The browser has no capabilities for data processing, so all data processing must be performed at the server.  This is potentially problematic in two instances.  First, the server might become overloaded processing multiple jobs, which could be more easily handled by the client. Second, when an image such as a chart is created, the data sent to the WWW browser is a graphic file which will be a much larger file than the actual data used to generate the image.  If the client uses data from the server to create an image, the amount of network traffic would be significantly reduced, leading to improved performance.  A final problem is that while WWW browsers support some types of servers without modifications to the server, database and WAIS servers require development work on the server to allow access.  Multiple options are available and documented for WAIS servers, and several examples exist of providing access to Oracle and other relational databases, but access to database and custom servers may involve significant server modification and development work.  In some instances, it may not be feasible to develop an interface from WWW because the server requires a stated protocol. Unfortunately, WWW browsers alone are not a universal solution for Internet server access.

However, the advantages of utilizing WWW browsers and servers in information system design should not be minimized. From the development perspective, WWW browsers exist for all major operating systems, and the development of clients can be anticipated to continue as new operating systems emerge.  This removes a major cost of information system development and support.  From the user perspective, WWW browsers provide access to a range of services and are thus more likely to be installed and used on a regular basis than custom clients for a particular information system.  This creates a major incentive for resource providers to utilize WWW servers since they are immediately available to the existing and massive install base of WWW browser users.

The development of applications such as WWW have facilitated the distribution of data and information resources over the Internet. While this has lead to an explosion of available resources, a specific resource may be often difficult to locate. A frequent quote overheard on the Internet is: "Everything you need to know is on the Internet. You just can't find it."[6] By design, the Internet is a cooperative, unmanaged venture. This is one reason why the Internet has been so successful, but it is also the source of the grand challenge of locating resources.

Several efforts have been undertaken to chart the Internet or at least provide mechanisms for facilitating the location of data and information resources. One listing of these efforts includes eighteen different searchable catalogs of Internet resources.[7] Some efforts focus on cataloging specific types of resources. For example, the Council of European Social Science Data Archives (CESSDA) is developing an interface to the distributed collections of European and other social science data archives. At present, this involves a global map of resources with pointers to individual archive resources.[8] The U.S. federal government has created the Government Information Locator Service (GILS)[9] to facilitate the cataloging government resources using a standard system. The G-7 countries have agreed to prototype the GILS standards for non-U.S. resources as well, but the GILS project is still very much in the early phases of development. Another effort is being undertaken by the Consortium for International Earth Science Information Network (CIESIN). CIESIN has developed the CIESIN Gateway as a system for searching distributed metadata collections and providing access to heterogeneous resource servers.[10] While these projects are still in development or early phases of implementation, the existing results show that the technology exists for solving the problem of locating resources on the Internet.

One of the approaches for locating data and information resources that has been implemented is the CIESIN Gateway. The CIESIN Gateway was initially developed to meet the requirements of CIESINÆs mission as the SEDAC (Socioeconomic Data and Applications Center) in NASA's Earth Observing System Data and Information System (EOSDIS). One of SEDAC's charges is to serve as a two-way gateway between social science and physical science researchers studying global change. The CIESIN Gateway fulfills this function by providing a single interface that allows searching of multiple data archives. The CIESIN Gateway includes a single interface that allows searching of the EOSDIS IMS, NASA Global Change Master Directory, and related directories of data, and also allows searching of key social science and other related metadata collections. Organizationally, CIESIN accomplishes this task through its Information Cooperative program. The Information Cooperative provides an institutional umbrella for linking together data centers worldwide. The CIESIN Gateway provides the technical implementation for the Information Cooperative by allowing searching of the directories of each of those data centers.

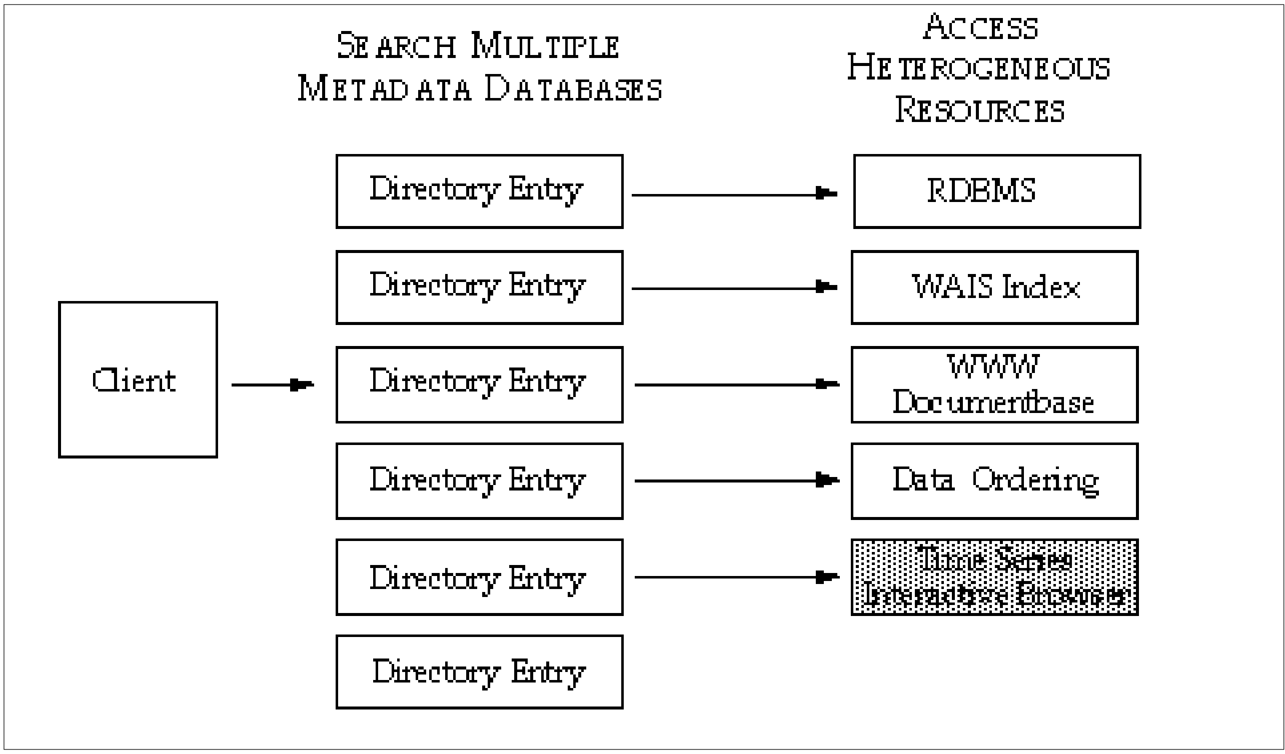


**Figure 1. CIESIN Gateway Conceptual Schematic**

The philosophy of the Information Cooperative is to encourage each partner data center to maintain their own metadata directories and data archives. This requires that the CIESIN Gateway be a distributed system, capable of searching multiple systems in parallel. Also, since each data center is likely to have different existing information systems, the CIESIN Gateway must support access to a variety of commercial databases and other Internet servers. In addition to searching and displaying high level metadata, the CIESIN Gateway also provides the capability of accessing on-line resources identified by the metadata. In some instances this is done within the CIESIN Gateway client, but in others it is accomplished by spawning an external application such as a WWW browser. A conceptual view of the capabilities provided by the CIESIN Gateway are described in Figure 1.

Thus, the CIESIN Gateway combines the capabilities of a search system for locating resources with capabilities for accessing those resources regardless of server type.

Based on the lessons learned from the development of the original CIESIN Gateway system, CIESIN, in collaboration with Brooklyn's Polytechnic University, is working on a new project, Raven. Raven subsumes the CIESIN Gateway functionality within a larger framework of access services. Raven presents the user with a list of services, such as CIESIN Gateway, WWW browser, and interfaces to custom applications and databases (see Figure 2).

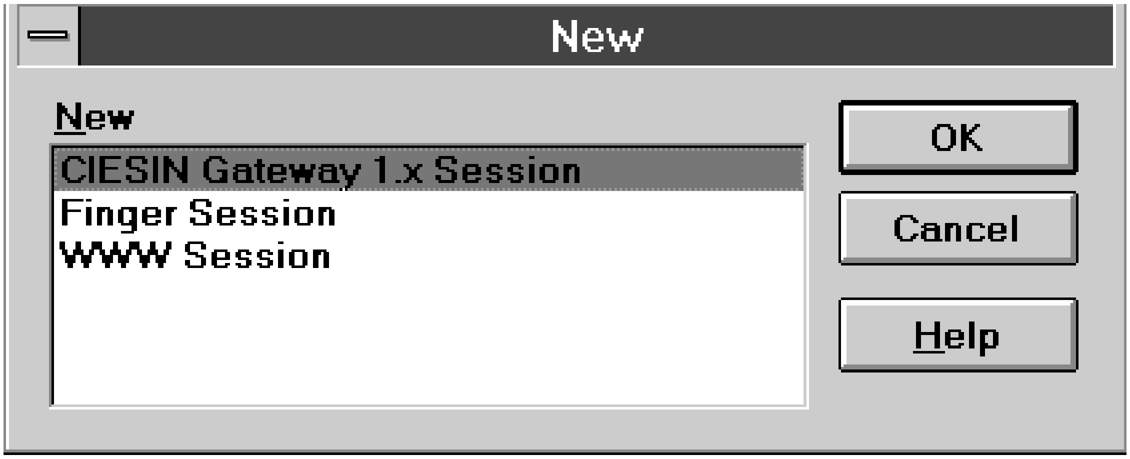
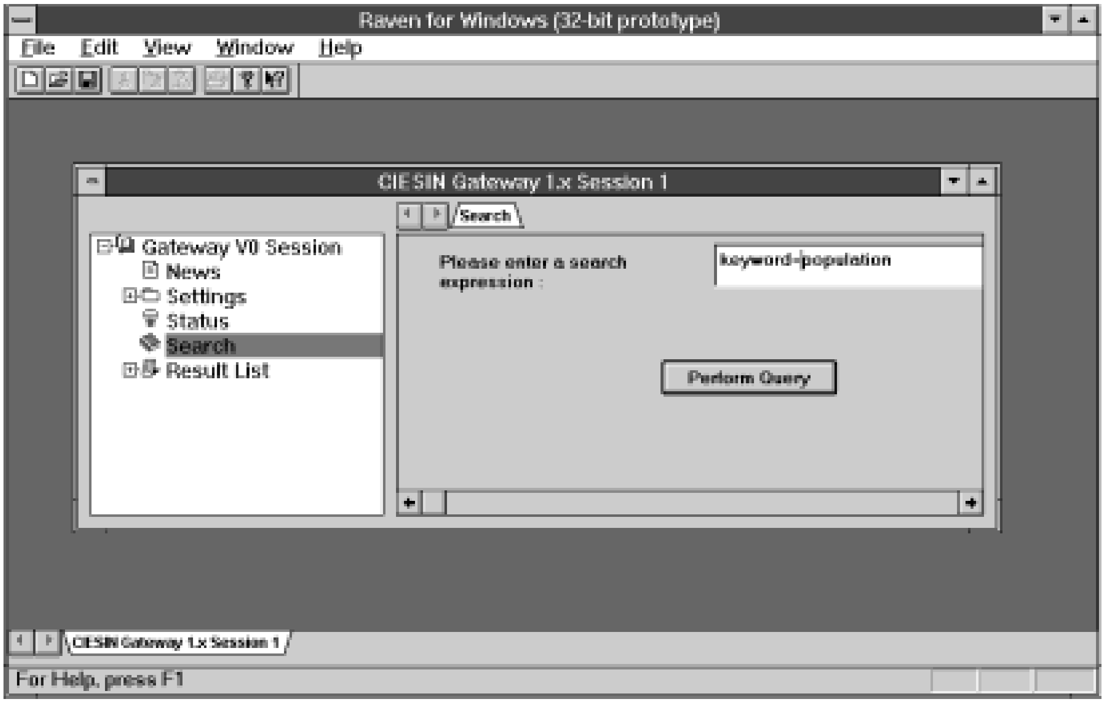
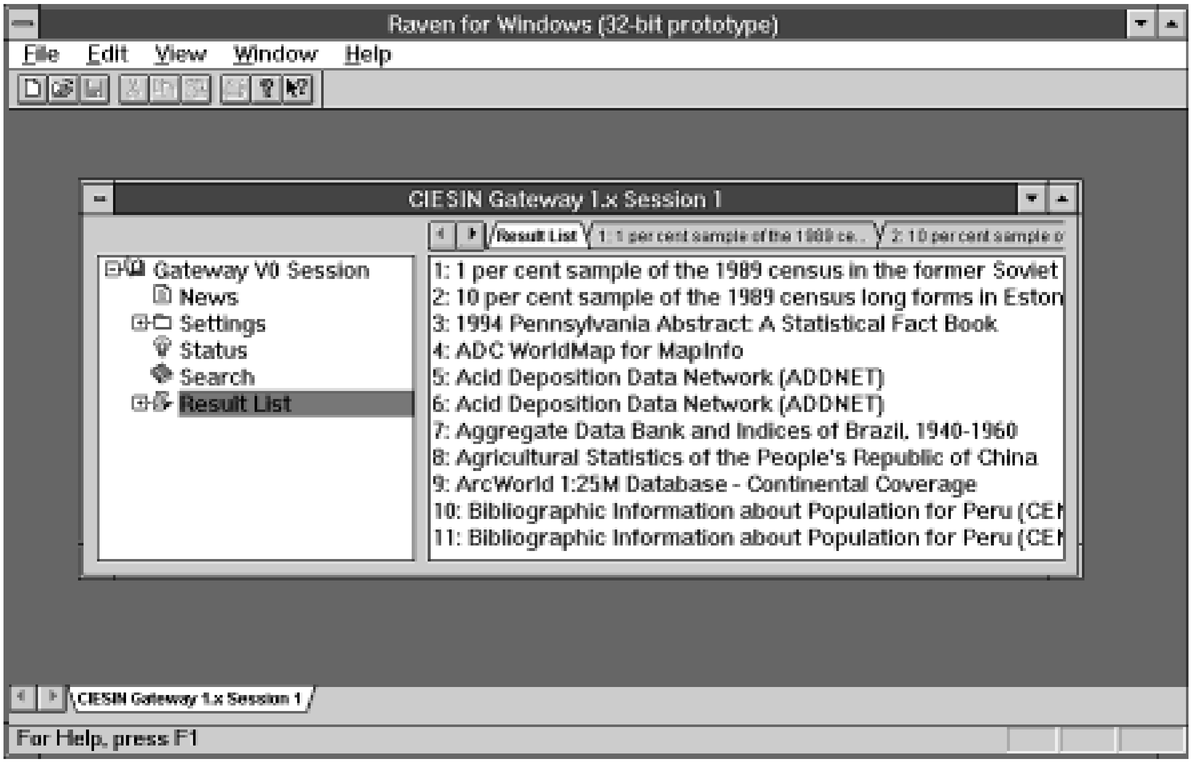


Figure 2. First Screen from Raven Prototype


Raven for Windows (32-bit prototype)
File Edit View Window Help
CIESIN Gateway 1.x Session 1
Search
Gateway V0 Session
News
Settings
Status
Search
Result List
Perform Query
CIESIN Gateway 1.x Session 1
For Help, press F1

Please enter a search expression : keyword=population

Figure 3. Search Screen from Raven Prototype


Raven for Windows (32-bit prototype)
File Edit View Window Help
CIESIN Gateway 1.x Session 1
Result List
1: 1 per cent sample of the 1989 ce..
2: 10 per cent sample o
Gateway V0 Session
News
Settings
Status
Search
Result List
CIESIN Gateway 1.x Session 1
For Help, press F1

1: 1 per cent sample of the 1989 census in the former Soviet
2: 10 per cent sample of the 1989 census long forms in Eston
3: 1994 Pennsylvania Abstract: A Statistical Fact Book
4: ADC WorldMap for MapInfo
5: Acid Deposition Data Network (ADDNET)
6: Acid Deposition Data Network (ADDNET)
7: Aggregate Data Bank and Indices of Brazil, 1940-1960
8: Agricultural Statistics of the People's Republic of China
9: ArcWorld 1:25M Database - Continental Coverage
10: Bibliographic Information about Population for Peru (CE
11: Bibliographic Information about Population for Peru (CE

Figure 4. Search Results Screen from Raven Prototype

This approach provides more functionality within one client, and it also enhances the interoperability between separate services. The Raven design and development is based on object oriented design techniques that allow new services to be easily built using common components from other services. At the most basic level this includes the networking component of the services, but it also includes capabilities for viewing tables and entering queries (see figures 3-4).

Raven also uses cross platform development tools to facilitate the development of versions for multiple operating systems. The goal of the Raven project is to create a client that provides a single interface to a vast array of resources and systems and provides information system developers with a set of tools for easily creating a user interface appropriate for a particular system.

Technology such as WWW and the CIESIN Gateway provide the technical capabilities that allow organizations to take advantage of the infrastructure of the Internet to distribute data and information resources. These capabilities can have several impacts on the organizations that utilize these technologies. First, the Internet infrastructure allows for the development of new products and services that allow the creation of digital libraries. Digital libraries include network accessible collections of data and information resources. Because these resources are available over the Internet, the physical location of the user and the library itself becomes irrelevant. Also, digital, on-line storage of resources allows enhanced tools for searching, accessing, and analyzing resources. Thus, data and information resource providers can utilize Internet technologies to expand their user base and the services they provide.

A more subtle but significant organizational impact of these technologies affects the very organization of data centers. An example of this is the structure of CIESIN's Information Cooperative. To the user, the Information Cooperative appears as a single archive, but in fact, it is a collection of archives linked through the Internet and the CIESIN Gateway. This organizational approach is a form of adhocracy, a term first coined by the futurist Alvin Toffler in his book Future Shock[11]. An adhocracy is based on small, specialized organizations that use information networks to coordinate their activities, and thus act like a larger organization. This approach can be more efficient and responsive than traditional hierarchical structures. This change is not limited to data and information resource providers. Similar changes are being experienced in a variety of industries, both public and private. As the global economy becomes increasingly information based and dependent, this pattern will increase in frequency. As with other types of organizations, two primary organizational roles emerge: providers and brokers[12]. Providers develop and disseminate resources and provide support on the use and understanding of those resources. Brokers work with providers to develop catalogs of resources across providers and to develop information systems to offer a common access system for users to the resources of multiple providers (see Figure 5).

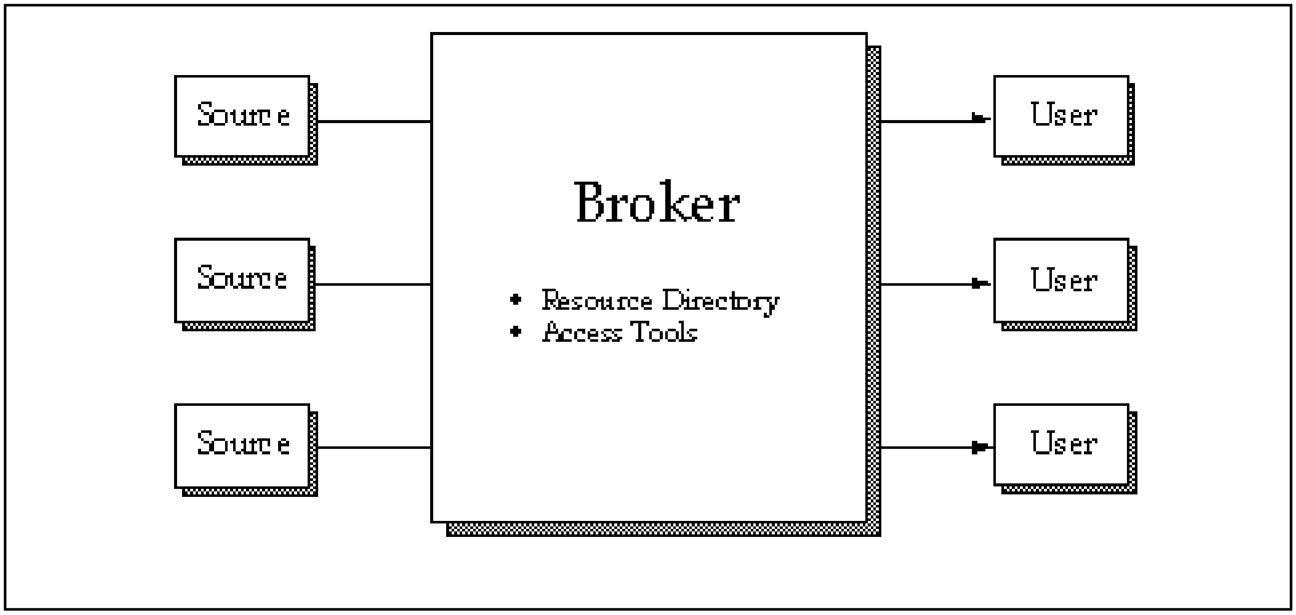


**Figure 5. Relationship between Providers, Brokers, and Users**

CIESIN, through its Information Cooperative program, represents one early effort at this approach. The CESSDA effort is another example. As the technical infrastructure for these types of organizational relationships becomes more widely implemented, other similar efforts will also emerge.

The Internet and related technologies have already had a significant impact on the practices of distributing and accessing data and information resources. Consider this closing thought from Christopher Locke (1994), "Unlike any previous medium, the Net's speed and reach seem to enable reaction to events that have not yet taken place. But this is an illusion. We are not seeing into the future, but more deeply into the present."[13] This capability will force data and information resource providers to continue to take advantage of new methods both of technology and organization to enhance and improve speedy and effective access and use of data and information.

1 Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.

2 The methods of user access for Database and Application servers are functionally equivalent.

3 Mailing lists are accessed through electronic mail and are functionally separate from any of the other client/server access / distribution methods.

4 The database interface is limited by the forms capability of the HTML standard and the existence of server side scripts for translating form input into a format understood by the database server and converting the results from the database and converting it to HTML.

5 A database front-end might be a separate client that runs on a users machine, or it might be a service that is accessed via telnet over the Internet. Generally, the front end is a custom interface to an off-the-shelf or custom server, usable only with a particular information system.

6 From David Lubar's "It's Not a Bug, It's a Feature": Computer Wit and Wisdom, Addison-Wesley Publishing, Reading, MA, 1995 who cites the source as "Anonymous, but common knowledge to anyone who's been there."

7 http://cuiwww.unige.ch/meta-index.html

[8] http://www.uib.no/nsd/diverse/untenland.htm

[9] http://www.usgs.gov/gils

[10] http://www.ciesin.org/gateway/gw-home.html

[11] Random House, New York, 1970.

[12] An excellent summary discussion on this is "Electronic Markets and Electronic Hierarchies," by Thomas W. Malone, Joanne Yates, and Robert I. Benjamin in Computer-Supported Cooperative Work: A Book of Readings, edited by Irene Grief, Morgan Kaufman Publishers, San Mateo, CA, 1988. Additional references are available from the author of this paper.

[13] From David Lubar's "It's Not a Bug, It's a Feature": Computer Wit and Wisdom, Addison-Wesley Publishing, Reading, MA, 1995.