ARTICLES

D-Lib Magazine
January 2001

Volume 7 Number 1

ISSN 1082-9873

Searching the Deep Web

Directed Query Engine Applications at the Department of Energy

Walter L. Warnick, PhD
U.S. Department of Energy
Office of Scientific and Technical Information
walter.warnick@science.doe.gov

Abe Lederman, MSCS
Innovative Web Applications
abe@iwapps.com

R.L. Scott, MPA
U.S. Department of Energy
Office of Scientific and Technical Information
scottrl@osti.gov

Karen J. Spence, MLS
U.S. Department of Energy
Office of Scientific and Technical Information
spencek@osti.gov

Lorrie A. Johnson, MSLIS
U.S. Department of Energy
Office of Scientific and Technical Information
johnsonl@osti.gov

Valerie S. Allen, MSLIS
U.S. Department of Energy
Office of Scientific and Technical Information
allenv@osti.gov

Abstract

Directed Query Engines, an emerging class of search engine specifically designed to access distributed resources on the deep web, offer the opportunity to create inexpensive digital libraries. Already, one such engine, Distributed Explorer, has been used to select and assemble high quality information resources and incorporate them into publicly available systems for the physical sciences. By nesting Directed Query Engines so that one query launches several other engines in a cascading fashion, enormous virtual collections may soon be assembled to form a comprehensive information infrastructure for the physical sciences. Once a Directed Query Engine has been configured for a set of information resources, distributed alerts tools can provide patrons with personalized, profile-based notices of recent additions to any of the selected resources. Due to the potentially enormous size and scope of Directed Query Engine applications, consideration must be given to issues surrounding the representation of large quantities of information from multiple, heterogeneous sources.

Introduction

While most search tools focus on the "surface web" or those web pages that are easily found and indexed by traditional web crawlers, the Department of Energy Office of

Scientific and Technical Information's (OSTI) interest has primarily focused on the "deep web." The term "deep web," or "invisible web," refers to a vast repository of underlying content, such as documents in online databases, that general-purpose web crawlers cannot reach. Both qualitative and quantitative in difference, deep web content is estimated at 500 times that of the surface web, yet has remained mostly untapped due to the limitations of traditional search engines [1].

OSTI, in partnership with Innovative Web Applications (IWA), released a product in April 1999 called EnergyPortal Search <http://www.osti.gov/energyportal> that addresses the search and retrieval of deep web content. The first product of its kind in government, EnergyPortal Search enables patrons to simultaneously search across distributed, deep web database content with a single search query.

IWA's novel Directed Query Engine, Distributed Explorer, has since served as the cornerstone for additional OSTI products and services requiring deep web searches. Each of these web products has applied Directed Query Engines in slightly different manners based on the unique attributes of site content, but in each case the deep web content is tapped and displayed.

By nesting Directed Query Engines so that one query launches several other search engines at host sites in a cascading fashion, the ability to assemble a comprehensive information infrastructure for the physical sciences is attainable. Simple algorithms plus enormous computing power can be a tremendous aid to human thought, thus benefiting the research and development process.

Background

Since 1947, OSTI's mission has been to collect, preserve, disseminate, and manage scientific and technical information (STI) for the Department of Energy (DOE). The DOE's scientific research community requires efficient and timely access to both current and legacy STI in order to carry out the research and development missions of the Department. Once relying primarily on paper-based processes, the manner in which OSTI's business is conducted has changed radically over the past several years with the successful deployment of digital technologies [2].

In September 1997, OSTI released the DOE Information Bridge, <http://www.osti.gov/bridge>, an online database currently containing over 60,000 searchable full text technical reports. This first version of the DOE Information Bridge was available only to DOE and DOE contractors. Soon after, in April 1998, a public version was made available in partnership with the Government Printing Office [3]. The DOE Information Bridge is part of the deep web, and it serves as the cornerstone of deep web Directed Query Engine searching at OSTI.

In addition to DOE's deep web collections such as the DOE Information Bridge and PubSCIENCE <http://www.osti.gov/pubscience>, many other federal agencies are adding to deep web STI offerings. Among these other agencies are the Department of Defense's Defense Technical Information Center (DTIC), which is rapidly increasing its STINET database to include full text [4]; the National Aeronautics and Space Administration (NASA) Technical Report Server [5]; the National Library of Medicine's PubMed [6], and other agencies as well. Such databases create huge deep web collections but are only useful if a patron knows where to find them.

The deep web will continue to grow as new full text material is added. In addition to full text information, the deep web may include images, presentations and other media that

are also essentially invisible to search engines. While OSTI has moved aggressively to digitize the gray literature of the Department of Energy, only approximately 12% has been made available electronically. The remaining 88% is in non-electronic formats and must be digitized for broad dissemination to be achieved. As other agencies also digitize their legacy collections, the deep web will further expand.

Directed Query Engine Searching Introduced

OSTI began building a virtual library collection of scientific and technical information in November 1997, using the DOE Information Bridge as a core collection. The EnergyFiles Virtual Library <http://www.osti.gov/energyfiles> includes not only the DOE Information Bridge, PubSCIENCE, and other OSTI products but links to over 400 energy-related collections and resources. Even as EnergyFiles was being compiled, however, it was recognized that mere collections do not make a library. Rather, a library requires that the information be organized and retrievable. Thus, it was determined that a search feature for EnergyFiles was necessary. Although the phrase "deep web" was not yet in the lexicon, it was immediately evident that common web search engines would not be capable of searching the rich full text of collections like the Information Bridge. Since the ability to search throughout the full text documents was the ultimate goal, alternatives were explored.

In its quest to enable patrons to search the deep web content of its own databases, along with external sources, OSTI collaborated with IWA, a small business in Los Alamos, New Mexico. Founded in 1996, IWA develops sophisticated web applications for aggregating, managing, and disseminating content for portals and other information networks. In collaboration with OSTI, IWA developed the proprietary Distributed Explorer Directed Query Engine, which supports several OSTI products used for deep web searching [7].

In conjunction with IWA, the OSTI Virtual Library Team identified eleven of the most popular databases from EnergyFiles. Multiple subject categories were represented. The Distributed Explorer Directed Query Engine was then configured to search these multiple heterogeneous databases in parallel, searching down to the document word level when this capability was available at the site itself. The searches are sent instantaneously and the information is retrieved within only a few seconds from around the country. The results are then displayed through a single interface that can be navigated from the patron's desktop. Thus, EnergyPortal Search was born.

To OSTI's knowledge, the EnergyPortal Search implementation in April 1999 was the first Directed Query Engine application in government which addressed the complexities of deep web searching [8]. At its core, it brought together and integrated the multiple heterogeneous web applications developed at OSTI. In addition, other DOE databases as well as popular databases developed by other government agencies such as the National Institutes of Health (NIH), were incorporated into the Directed Query Engine search.

By using this innovative technology, it no longer mattered where the information resided nor what format it was in, and the patron did not need to know which agency posted the information. These factors no longer posed barriers to the process of information discovery. With the success of this project apparent, fifteen additional databases were integrated into EnergyPortal Search. All can be searched with a single directed query.

The tremendous power of Directed Query Engine searching can be demonstrated by the ease by which vast quantities of information are made retrievable:

- The research, which underpins the growing 62,000 full text technical reports and 5 million pages residing in the DOE Information Bridge, roughly required 100,000 man-years to conduct at the various DOE labs and facilities;
- The reports that integrate the acquired data took roughly 1,000 man-years to write;
- The DOE Information Bridge database itself took approximately 20 man-years to build;
- To bring the DOE Information Bridge together with other equally extensive resources, making them all searchable through one single-query interface, took only man-days of effort, not man-years.

Building on Directed Query Engine Success

Building upon the successful implementation of Directed Query Engine searching with the EnergyPortal Search product, OSTI and IWA began to investigate other ways in which this break-through technology could be used to uncover the vast amounts of scientific and technical information contained within the deep web. Several additional products based on Directed Query Engine searching have been released during the past year.

The PrePRINT Network, <http://www.osti.gov/preprints> released by OSTI in January 2000, introduced the use of a Directed Query Engine to provide fielded distributed searching specific to a particular information type. In this case, the deep web content of preprint servers and databases has been made readily available to patrons through a single search interface. The PrePRINT Network is a searchable gateway to web-based collections of scientific preprints and reprints as provided by the scientists themselves. Preprints, or "e-prints," are manuscripts that have not yet been published, but may have been reviewed and accepted; submitted for publication; or intended for publication and being circulated for comment. The preprint collections represented in the PrePRINT Network are provided by a variety of sources, including the researchers themselves or their institutions. The content of the individual preprints, as well as the technology used by the various databases and servers, remains within the purview of the site owners and requires no special formatting or standardization in order to be included in the PrePRINT Network. As with other products employing the distributed searching capabilities of a Directed Query Engine, the PrePRINT Network provides the ability to simultaneously search across multiple preprint databases and servers.

Recently implemented, the PrePRINT Alerts <http://www.osti.gov/preprintalerts> feature provides patrons with personalized, profile-based notices of recent additions to any of the selected resources. Patrons simply set up a profile defining their interests, and as new information is added to the preprint servers selected, the patron is notified via email. Building on the initial configuration of the Directed Query Engine for preprints, the Alerts service was designed specifically to meet the needs of researchers who desire to control the volume and content of the data they receive.

Two new tools have been developed at OSTI over the past several months using the Distributed Explorer Directed Query Engine to search and combine different sets of Federal information. Developed in response to the Physical Sciences Information Infrastructure (PSII) Workshop, <http://www.osti.gov/physicalsciences> held in May 2000, these two sites signify new collaboration among Federal agencies to enable convenient access to government information by the American public. With these tools, it is no longer necessary for a patron to know which agency is working in a particular area

or discipline. They support an interdisciplinary view of science by providing the opportunity to look beyond one's specialization and to access and combine relevant information from other disciplines.

The first tool, the GrayLIT Network <http://www.osti.gov/graylit> was created as a result of an agreement in principle between the Department of Energy, the Department of Defense - Defense Technical Information Center (DTIC), NASA and later the Environmental Protection Agency (EPA). Each of these agencies had already developed one or more extensive electronic collections of full text gray literature freely available at their agency's web sites. The GrayLIT Network provides a single interface to search across the combined deep web information without placing any burden in the participating agencies. It met the call by Workshop participants for an early success (Figure 1).

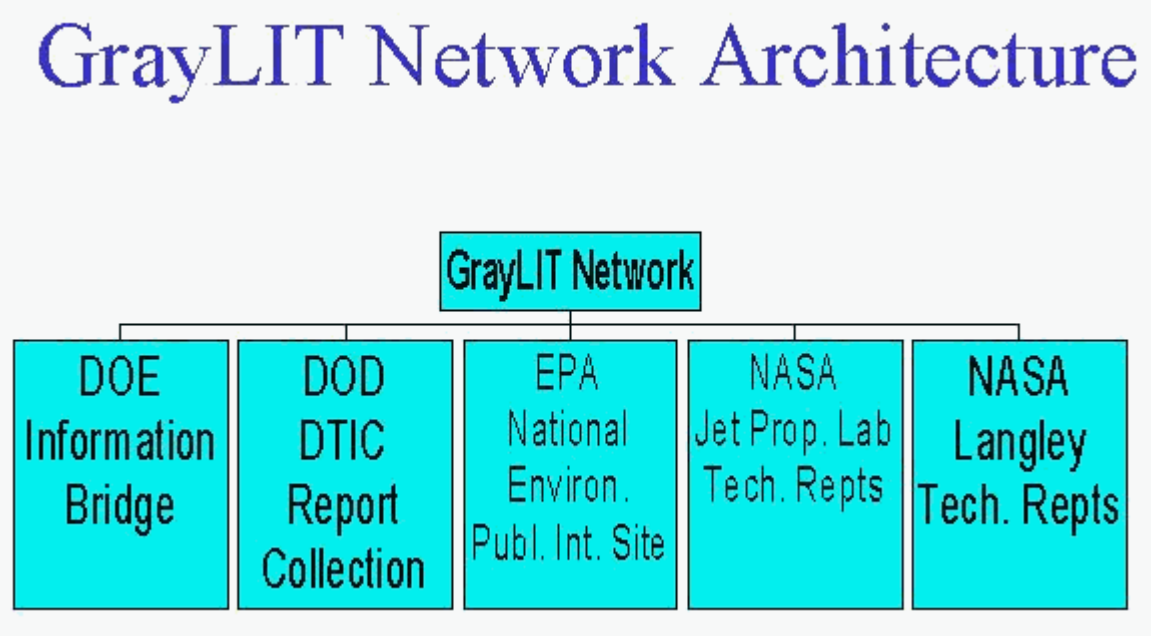

Figure 1 - GrayLIT Network Architecture

The second new tool, Federal R&D Project Summaries, <http://www.osti.gov/fedrnd> combines over 240,000 records of research and development summary and awards data from three governmental agencies: DOE, NIH and the National Science Foundation (NSF). The public may use this tool to stay better informed about taxpayer investments in research and development. Federal R&D Project Summaries provides a unique window to the Federal research community, allowing agencies to better understand the research and development efforts of their counterparts in government, and it enhances the potential of symbiotic research activities. Like the GrayLIT Network, Federal R&D Project Summaries places no burden on the participating agencies while meeting the call of the PSII Workshop for early successes (Figure 2).

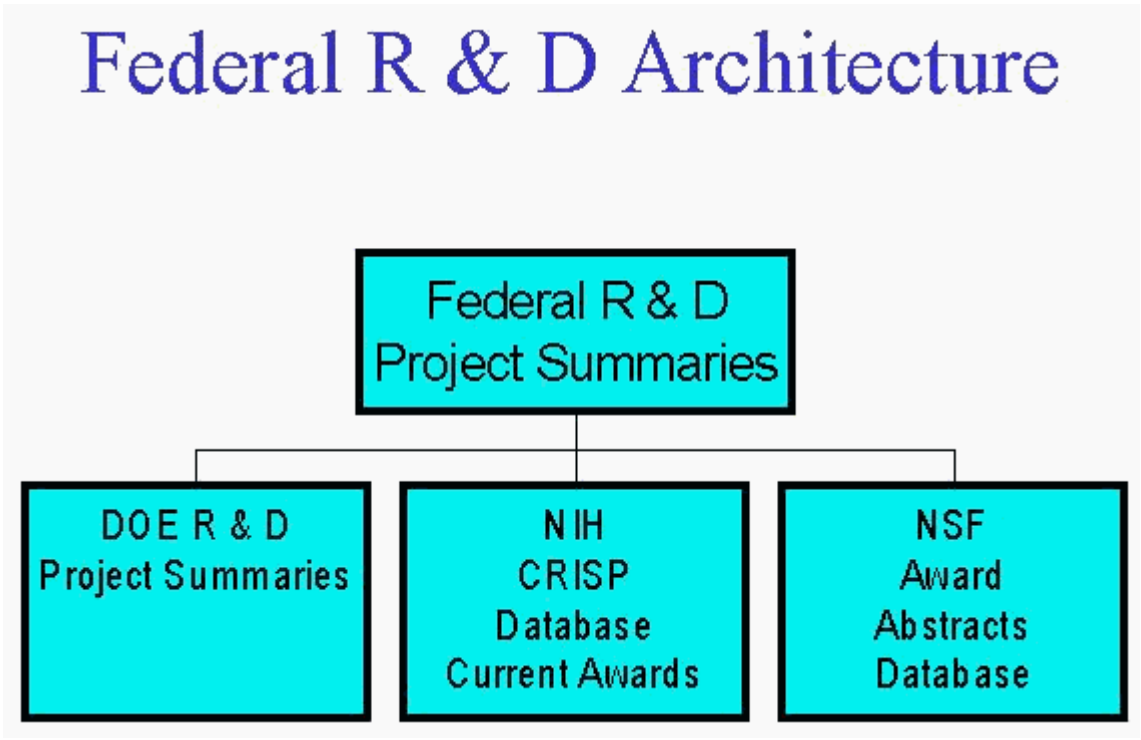

Figure 2 - Federal R&D Architecture

Features and Benefits of the Directed Query Engine to Information Patrons and Creators

Features of the Distributed Explorer Directed Query Engine include the capability to Mark and Download specific records, allowing patrons to save and manipulate their search results sets. Distributed Explorer also allows the patron to create printable versions of marked documents, create tagged records from marked documents for downloading, and build searchable, personal online libraries of marked documents.

The primary significance of the Distributed Explorer Directed Query Engine approach used with these novel web products is that a single search can open the door to a wealth of information from the vast virtual collections that make up the realm of the deep web. The patron does not need to know which database or source might contain the needed information, nor does he/she need to know the information format.

The deep web includes many vast collections, which may be viewed as the building blocks of digital libraries. The Distributed Explorer Directed Query Engine searching capabilities serve as the glue holding the building blocks together. As Directed Query Engines launch simultaneous searches onto geographically dispersed servers, the patron gains a power comparable to that of parallel processing.

Directed Query Engine searching also provides benefits to the creators and originators of information. Since Directed Query Engines work by launching searches on the engines residing at each individual source, no obligations are placed on site owners to change their current processes or configurations. The Directed Query Engine is configured to work with the content and with the search and retrieval capabilities offered by each target source or database. With no requirements for metadata or other standardizations, Directed Query Engine searching also places no additional burdens on information creators. While

such initiatives as the Open Archives Initiative have been popular recently, they do mandate a standardized protocol for transmitting and querying of metadata [9]. Directed Query Engine searching is not in competition with such initiatives; rather if these initiatives are successful, the Directed Query Engine approach will work even better. For example, common metadata standards would allow for more precise fielded searching within the target databases accessed by the Directed Query Engine.

How Distributed Explorer Works

A cascading hierarchy of IWA's Distributed Explorer applications is the core technology that supports nested Directed Query Engines, and it is the facility with which Distributed Explorer applications can be nested and deployed that enables directed queries against hundreds -- even thousands -- of deep web databases.

Basically, Distributed Explorer can search any online database that has a web interface, including Z39.50 databases. Distributed Explorer relies on a series of small search configuration files and user interface files. The search configuration files contain the instructions for interacting with each online database used by an application (e.g., there is one configuration file for each preprint database contained in the PrePRINT Network). Each time a new database or other source is added to a Distributed Explorer application, the developers use special software tools to create the base file, and then customize the search configuration to handle exceptions and any special requirements of the host site. At regular intervals, Distributed Explorer automatically checks the databases and other sources to see if changes need to be made to search configuration files.

A patron initiates a directed query by selecting the sources to be searched and entering the word, phrase, and operators that constitute the query. Immediately, Distributed Explorer reads the search configuration file for each selected source, translates the query into the syntax expected by the respective sources, and submits the query in parallel to each selected source. Fielded searches against databases where only a subset of fields are available for searching are handled by defaulting to the most specific field available and combining the remaining search fields with a general full-text search of the database. While a directed query is in progress, Distributed Explorer monitors and reports out the status of the query at each selected site or database. Once the result sets are complete, Distributed Explorer reads the user interface configuration file for the host site, then formats and displays the results from each selected source according to the instructions contained in this file.

The user interface configuration files contain the parameters and instructions that Distributed Explorer uses to format and display results from each queried database with the user interface specifications of the host site. Typically, there is considerable variation in the way individual databases display search results on their respective sites, so displaying results as-is would limit the usefulness of results from multiple directed queries. To overcome this handicap, Distributed Explorer parses each set of results and reformats the display using the specifications contained in the user interface configuration files for the host site. In addition to providing the look-and-feel of the host site, Distributed Explorer also adds search term highlights and navigation features (e.g., next and previous document) on the results list and on each document display.

Implementing Distributed Explorer for Nested Directed Query Engine Searches

As a potential next step, it is possible to nest Directed Query Engine applications. Such an architecture offers the advantages that the existing search configuration can be

leveraged, and the user interface files can be used for each Distributed Explorer application selected for the hierarchy (Figure 3). Though a particular Distributed Explorer Directed Query Engine application may provide search access to dozens of databases and other sources (which need not be arranged by product type or discipline area), only one search configuration file is necessary to include this application and all of its sources in a hierarchy of nested Directed Query Engine applications. This architecture would facilitate the search of large groups of sources.

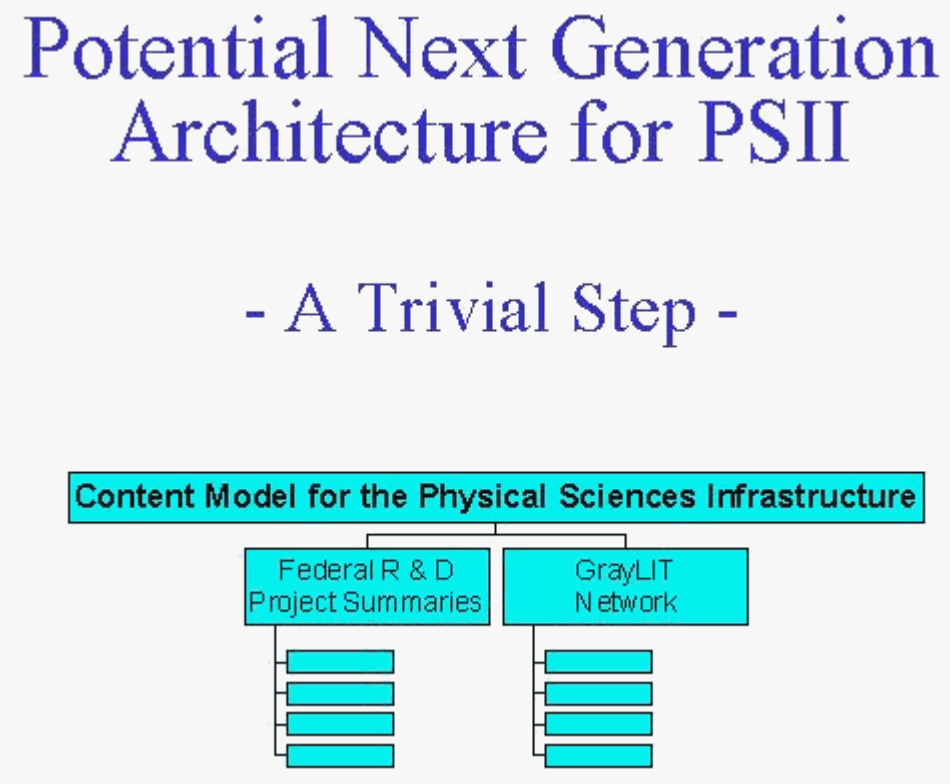

Figure 3 - Potential Next Generation Architecture for PSII

Given the potential breadth of sources from which patrons may choose during the course of a search session, user interface issues have arisen with respect to how to meaningfully present vast quantities of information. Currently, the results of a particular query are organized on the screen by source, with a linked title and brief description of each item listed in the order provided by the original source. Though this method is effective for results from as many as ten distinct sources, it is somewhat cumbersome when more sources are added. To overcome this problem, IWA has been investigating clustering algorithms that will make long lists of results easier to browse by grouping documents together based on content rather than source. This can help in locating interesting documents more easily and in getting an overview of the retrieved document set.

Guiding patrons to select the most relevant sources for searching in their subject area can be addressed by creating multi-term definitions and sets of keywords that describe the content and purview of each database or other source. Applications can be configured to suggest sources that, though not selected by the patron, might correspond to the search terms being submitted. This will become an increasingly important tool for patrons as the size and scope of Distributed Query Engine applications grow.

Future Directions

Many organizations have site-wide or individual subscriptions to online databases that require patrons to provide institutional or individual passwords to gain access. In the near future, IWA will be enhancing Distributed Explorer to manage access to multiple, password-protected sources for each patron. Using secure authentication protocols and secure personal account records containing access information for individual patrons, Distributed Explorer will require but a single password and login to unlock access for directed queries against any database for which the patron has privileges.

Clustering algorithms will continue to be developed to improve the usability of long results lists. IWA is also working on several tools for generating graphical versions of these lists. With an intuitive, point-and-click interface -- instead of long lists of sources and titles -- patrons will be able to visually navigate and browse their search results by cluster, source, or relationship.

Another planned enhancement to the Distributed Explorer Distributed Query Engine will provide dynamic filtering and cross-linking of the contents of documents from various sources. For example, if an abstract contains a reference to a source journal and article that are available elsewhere on the web, Distributed Explorer will create a direct link to the article using either a centralized DOI-based (Digital Object Identifier) locator or a customized locator database until DOI is widely implemented. In addition to supporting DOI, XML, and other document sharing initiatives, IWA will continue to monitor the growth and evolution of Z39.50 databases. When Z39.50-compliant databases become more numerous, a separate Z39.50 module will be implemented for Distributed Explorer that will provide enhanced performance and access.

The architecture outlined above should be considered as only the first generation for the Physical Sciences Information Infrastructure. Besides harnessing vast amounts of distributed information and making it accessible to a huge distributed audience, the architecture has the essential advantage of being easily and inexpensively deployed. There still remains much more to be done to obtain even greater utility for accessing physical sciences information requiring next generation architectures.

The capability to access and collect information automatically from a huge number of Web sources exists today. The ability to index, or map this information to a searchable database, provide relevance ranking of search results, and provide access to the intellectual content of those records without violating copyright can also be done today. However, new efforts must be focused on the accuracy and relevance of the information retrieved through deep web searching.

Key to improving relevance is the use of controlled vocabulary software programs that map and count the frequency of terms consistent with the subject category of interest. While the use of controlled vocabulary may seem to some a regressive approach to information science, the advanced state of computer technology now allows the opportunity to capitalize on the tremendous intellectual investment that has already been made in STI thesauri. Running selected subject categories as an overlay software on the results of a deep web search can provide the patron more accurate and focused search results. This second-generation approach could provide the necessary tool set to make deep web searching of greater value than heretofore considered possible.

Conclusion

In today's environment of increased expectations, information consumers demand tools and capabilities that help them access relevant information and enable them to manipulate the information that they retrieve. As the deep web continues to grow deeper, the

importance of Directed Query Engine tools increases proportionately and the Distributed Explorer application is well positioned as the first generation architecture for the PSII. The vision behind the PSII is to create an integrated network for the physical sciences where content, technology and service converge to make resources readily accessible, openly available, useful, and usable. There is no value in information that is inaccessible, but great value in relevant information accessible with just a few clicks and commands.

References

1. BrightPlanet.com LLC. "The Deep Web: Surfacing Hidden Value." White Paper, July 2000. Available <http://completeplanet.com/Tutorials/DeepWeb/index.asp>

2. U.S. Department of Energy Office of Scientific and Technical Information. December 2000. Available <http://www.osti.gov>

3. U.S. Government Printing Office. October 2000. Available <http://www.access.gpo.gov/su_docs>

4. U.S. Department of Defense, Defense Technical Information Center (DTIC), Scientific and Technical Information Network (STINET). October 2000. Available <http://stinet.dtic.mil>

5. U.S. National Aeronautics and Space Administration Technical Report Server. July 31, 2000. Available <http://techreports.larc.nasa.gov/cgi-bin/NTRS

6. National Library of Medicine PubMed. December 2000. Available <http://www.ncbi.nlm.nih.gov/PubMed/>

7. Innovative Web Applications. December 2000. Available <http://www.iwapps.com>

8. Varon, Elana. "DOE Energizes Site with Extensive Web Searches." *Federal Computer Week*. August 16, 1999. Available <http://www.fcw.com/fcw/articles/1999/FCW_081699_956.asp>

9. The Open Archives Initiative. November 21, 2000. Available <http://www.openarchives.org>

Top | Contents
Search | Author Index | Title Index | Back Issues
Previous Article | In Brief
Home | E-mail the Editor

D-Lib Magazine Access Terms and Conditions

DOI: 10.1045/january2001-warnick