Searching the Deep Web – Distributed Explorit Directed Query Applications

Valerie S. Allen, MSLIS
U.S. Department of Energy (DOE)
Office of Scientific and Technical Information
Oak Ridge, Tennessee 37830
1.865-576-3469
allenv@osti.gov

Abe Lederman, MSCS
Innovative Web Applications (IWA)
134 East Gate Drive
Los Alamos, New Mexico 87544
1.505-663-1302
abe@iwa-solutions.com

Abstract: In 1999 a directed query distributed search engine was integrated into a new Department of Energy Virtual Library of Energy Science and Technology. Millions of pages of government information across multiple agencies were made immediately searchable via one query, setting the stage for the development of a variety of interagency initiatives and applications.

Categories and Subject Descriptors: Digital Libraries

General Terms: Design, Experimentation

Keywords: Deep web, distributed searching, directed query, Federal agencies, cross-agency portals, alert system

Searching the Deep Web – Distributed Explorit Directed Query Engine Applications

Deep web content is estimated at 500 times that of the surface web, yet has remained mostly untapped due to the limited capabilities of common search engines [1]. Federal Agencies and others working with scholarly information are concerned with the mechanisms of making this information available to their user groups through the provision of effective search and access options.

To this end, the Department of Energy (DOE) Office of Scientific and Technical Information (OSTI) formed an alliance with Innovative Web Applications (IWA) in Los Alamos, New Mexico, and together the two have been developing sophisticated web applications for aggregating, managing, and disseminating content [2]. The Distributed Explorit Directed Query Engine, tested and implemented in several DOE websites since 1998, relies on small search configuration files and user interface files. The search configuration files contain instructions for interacting with each online database used by an application. Distributed Explorit can search any online database that has a web interface, including but not limited to Z39.50 databases. Each time a new database or other source is added to a Distributed Explorit application, the developers use special software tools to create the base search configuration file and then customize this file to handle exceptions and any special requirements of the host site. Automated tools monitor 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 Explorit 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. When performing fielded searches against a number of sources, Distributed Explorit intelligently maps fields being searched against available fields. While a directed query is in progress, Distributed Explorit monitors the status of the execution of the query at each selected site and, as soon as each result is returned, immediately displays the results to the user. Fields are extracted from a results list based on rules in the search configuration file, then displayed in a uniform manner, based on the specifications in the user interface configuration file.

The user interface configuration files contain the parameters and instructions that Distributed Explorit uses to format and display results from each queried database with the user interface specifications of the host site. Distributed Explorit parses each set of results and reformats the display using the specifications contained in the user interface configuration files for the host site. Building on the technology of Distributed Explorit, the Explorit Alerts feature provides patrons with personalized, profile-based notices of recent additions to any of their selected resources. These tools will be demonstrated within the framework of Department of Energy operational and development systems: GrayLIT Network (http://www.osti.gov/graylit); Federal R&D Project Summaries (http://www.osti.gov/fedrnd); PrePRINT Network (http:/www.osti.gov/preprint); EnergyFiles Virtual Library (http://www.osti.gov/energyfiles); and Science.gov, under development (http://www.science.gov).

SIGIR'01, September 9-12, 2001, New Orleans, Louisiana, USA.
ACM 1-58113-331-6/01/0009.

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

2. 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