How to Maximize Your Strategic Investment in Federated Search

Whitepaper prepared by Deep Web Technologies

January 22, 2007

How to Maximize Your Strategic Investment in Federated Search

Research "time-to-answer" success depends on accelerating information discovery
The research process is complicated, time-consuming, and costly, and success results from getting to the answers first. Globalization of markets, competition, and regulations demand superior enterprise collaboration and intensify the need to accelerate multi-discipline analysis and discovery. It is impractical for researchers to spend time serially searching and sifting through hundreds or thousands of information sources in various disciplines. Automating repetitive information search steps allows specialists to maximize time spent performing the analysis they are trained to do in pursuit of answers. An important capability to speeding up the information access component of research is federated search. Multiple sources are accessed and searched simultaneously. Source and result analytics help identification, sharing and reuse of the best information, to support answer-finding objectives.

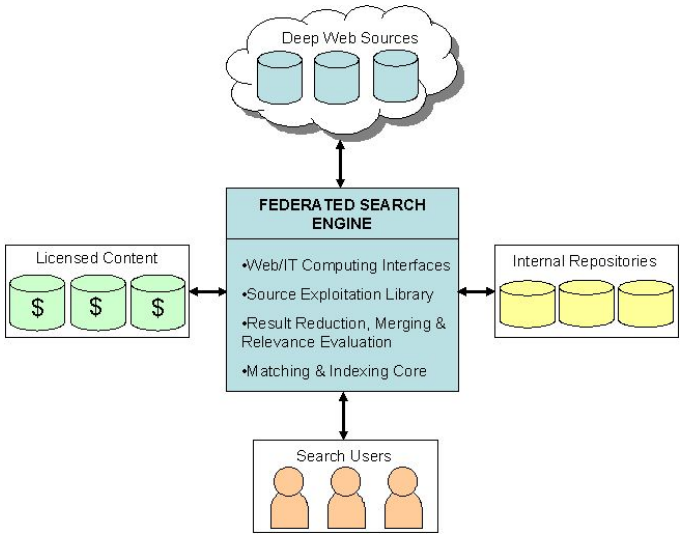
Well implemented federated search will help tame a chaotic information environment where:

- Hundreds or maybe thousands of data bases where the best sources are "thrown in" with the worst
- No clear path to begin your analysis; no obvious "winners" appear early
- Expectations of "Google-like" simplicity are not met and users give up before finding what that really need for informed decisions

Federated search must extract the best each information source has to offer and meet ease-of-use and result relevance expectations of users
For the specific strategic advantage an organization needs, a federated search solution must exploit:

- The organization's subject areas, reasons, questions, purposes, and information sources
- The organization's computing and search/research workflow environment

Federated Search Environment for Information Discovery



Automate researcher finding and sifting to maximize time for analysis

To achieve the two objectives above, an effective federated search solution provides the following functions:

1. Information Source Exploitation
2. Search User interface
3. Retrieved Results Analysis & Presentation
4. Web/ IT Environment Integration
5. Scalability
6. Customizations

Source Exploitation

The most important component of federated search is the exploitation of information sources. Source exploitation is the "knowledge base" of federated search. Information sources come with a wide variety of access characteristics, search engine sophistication, interface complexity and capabilities to parse results. Since the federated search objective is to use one interface to simultaneously search many sources and generate one amalgamated result set, the source exploitation function must be able to extract and normalize sources:

- Sources on the Internet
  - Surface Web
  - Deep Web
- Proprietary internal sources/collections
- Subscription databases

To ensure that a search retrieves the most relevant documents for a user's search request, all source connectors must be configured to access the best each information source has to offer. To obtain results, parse result pages to extract specific titles, URLs, authors, abstracts, dates, snippet and full text, and present results from multiple sources, federated search should exploit the whole population of sources in all of the following ways:

- "Screen scraping", the ability to search a source through its HTTP interface. The majority of information sources are only searchable through this approach. Managing cookie-based sessions or sessions embedded in search pages and result pages.
- XML Gateways. A number of content providers are now providing XML-based interfaces specifically designed for use by federated search
- Web Service Interface. SR/W (Search Retrieve via Web Services) based on the library Z39.50 standard is emerging as a new standard for access to content repositories
- Custom connection to Content Management Systems, E-mail systems, etc
- Licensed content, submission of individual user credentials as part of every search request; single sign-on environments

Leaving a connection method out will simply degrade the value of information sources that rely on it for native search.

The Search User Interface

When sources are exploited to the maximum extent possible, the next step is to get the most out of each federated search request. Since every strategic federated search application is unique to its organization, the user interface is really a set of tools to be integrated into the organizations Web-IT interface accordingly.

- Source selection by subject-search area, drive searches by subject to "better", more reliable, more credible, sources, not to rule out outliers, but to focus valuable time on sources with known scope and credentials
- Subject Area taxonomy-ontology integration

- Saved, Repetitive and Subject Monitoring Searches, the latter with frequency controls and alerting mechanisms
- Search sharing
- Search expressions that enable pattern matching logic including
    - Taxonomy-ontology entity selection
    - Boolean logic, wildcard and thesaurus
    - Searching by field/location (source, author, title, date/range, full text)
    - Proximity
    - Combinations of the above

Results Analysis & Presentation

Search results analysis and presentation cannot substitute for effective searches, but they can dramatically speed up the answer finding process. Presented results are individual records-entries-documents which pass some source and/or search request criteria. Federated search is frequently used as a locator which serves a different function than an enterprise search engine with a single index and collection. Federated search results analysis and presentation must optimize the location process and enable a smooth transition to full content analytics. As with the search user interface, the capabilities below are most effective when provided as enhancement tools that are uniquely incorporated into the organization's computing environment.

- Incremental display of search results, allowing review to start while searching continues in the background. This capability's effectiveness is tied to the proper source exploitation and search requirement listed above, as well as relevance determinations
- Result list sorting and navigation
- Options to present relevant search results independent of source, as well as in the order returned by each individual source
- Retrieval of next sets of results
- Results sorted by date, title, author, etc. in addition to rank
- Capture/saving of results
- Interface to full content if result is only an abstract
- Incorporation of result content into enterprise content
- Ranking & Relevance

In structured data, relevance means "exact match", such as a date or name match. In unstructured text, relevance means "fit" between a chunk of text and a specific issue or subject as it is represented in a search request. Federated search must accommodate both types of content. At all times, a substantial component of true relevance is in the eye of the beholder, the mind of the researcher.

Federated search should capture practical, justifiable relevance as described below.

- Results from a "good" source, as described in the "search" section
- Results that "contained" the search request, in title, snippet, abstract, full-text etc.. "Containing" is subject to the characteristics of taxonomy elements, or the search expression language.
- Content previously viewed and determined to be relevant by the current user or a user in his/her "network"

Again relevance determination methods must be suited to the organization and its user requirements, and at least enable tools that support other helpful capabilities.

- Explicit ranking based upon search expression appearance in title, snippet , abstract and full-text
- Visualization, clustering and statistical "nearness" of result to search expression

Web/ IT Environment Integration

For maximum utility, federated search must be tightly integrated into the vertical-organization IT/web/portal environment. The starting point for achieving a tight integration is via an application program interface toolkit that separates the user interface layer from the back-end engine.

The support of open standards as defined by the grid community, OASIS, W3C, and service-oriented architectures further enables customers to maximize their existing architectures and functionality, including a tight integration with document management and workflow functions.

Scalability

Federated search scalability dimensions are:

- Sources
- Source changes
- Number of simultaneous searches/ ability to distribute computing resource
- Number of unique users

Any limitation to the use of viable sources is too critical to be allowed in strategic applications. Usage increases and their impact on performance ultimately drive the user experience, and the ability of the organization's architecture to accommodate federated search.

Customizations

The incredible revolution of the web over the last 10 years has shown distinct shifts in demand for Web 2.0 functionality. Real strategic advantage comes from an intense dedication and ability to adapt in near real time, and to extend the frontier of the application and technology. Federated search applications will last for a number of years, and will require upgrades. The scope of federated search applications, technologies and architectures is a substantial undertaking itself, and is best handled by organizations dedicated to the field.

Leading research organization uses a powerful federated search core and adapts to changing needs

Eight years ago the U.S. Department of Energy, Office of Scientific & Technical Information (OSTI) embarked on the following mission, "To accelerate discovery, it is essential to accelerate the diffusion of science knowledge". OSTI proposed the Department of Energy Science Accelerator, justifying the importance of its development by saying it is impractical for researchers to spend time finding and sifting through hundreds, if not thousands of information sources in various disciplines and still have time for life-altering discoveries of their own.

To execute on the mission, OSTI selected and partnered with Deep Web Technologies. Together they built the solution by starting with a robust and sophisticated search platform, and adding key enhancements from a comprehensive tool box of capabilities. They continue to work closely to achieve additional strategic advantage through customizations from its tool box.

A significant milestone was achieved in 2002 when http://www.science.gov introduced the capability to a consortium of 12 federal departments to search 30 major databases of federal science agencies. Other high profile solutions from this initiative include http://www.osti.gov/eprints and http://www.osti.gov/ScienceConferences

OSTI encouraged the scientific community to dispense with the popular misconception that Web technology is mature and that the Web already provides easy access to all meaningful information. In this spirit, OSTI pioneered federal government Web 2.0 applications for the public several years before the term was coined. It believes making science resources in the deep Web globally searchable cries out for a Web 2.0 solution and it is answering the call with the DOE Science Accelerator.

OSTI's development objectives remain aggressive and include the following highlights.

- Distributed access to all DOE's educational resources
- Improve digital access to DOE's pre-1990 R&D literature
- Enhance precision searching
- Develop next-generation relevance ranking algorithms
- Develop prototypes for analytical tools and grid computing
- Encourage collaboration through expanded services to share interactive technologies such as RSS, blogs, tags, and podcasts
- Overcome barriers to searching thousands of data bases in parallel
- Automate translation of queries and results
- Integrate text and numeric data

OSTI and the Department of Energy have invested more than $1M in R&D funding to achieve their unique research needs. Fortune 500 companies are also leveraging this capability by investigating and implementing the customizable platform made possible by this investment.

Prior to founding Deep Web Technologies in 2002, Abe Lederman was one of the founders of Verity where he helped develop one of the first text retrieval applications. He also founded Innovative Web Applications (now Doxcelerate) in 2000. Deep Web Technologies is dedicated to the development of cutting edge search and retrieval software solutions for government agencies and the private sector. In addition to OSTI successes, it has recently deployed a federated search solution for a Fortune 50 Technology Company that is seamlessly integrated with their corporate portal, including their in-house user interface. Deep Web Technologies is providing federated search functionality through a W3C compliant Web Services architecture which will be used to provision additional services to meet new requirements.

Know when to get help to achieve strategic research advantage from Federated Search
Deep Web Technologies Inc. has developed a "45 minute federated search for strategic advantage consultation" for professional research organizations which can be conducted over the phone. Outcomes include.

- Identify the # and types of databases to be searched
- Determine application & source security and authentication needs
- Specify result relevance requirement
    - # of search pages
    - source groupings on search pages
    - advanced search
- Define usage levels and hardware implications
- Outline connector creation and maintenance needs

The 45 minute session is conducted by the President/Chief Technology Officer of Deep Web Technologies Inc, Mr. Abe Lederman, who spearheaded federated search applications in 1999 for science and technology centers within the Department of Energy. This is not a sales presentation. It will consist of the best federated search requirements analysis and intelligence Mr. Lederman can supply in a 45 minute time-span. There is no charge for the call, but please be advised that the call must be strictly limited to 45 minutes.

The consult can typically take place within 1-2 weeks of your call. To secure a scheduled time, please call Darcy Pedersen 505.820.0301 or email her at darcy@deepwebtech.com and she will advise you regarding available time slots. She will also provide you with a pre-consultation questionnaire that will prepare both you and us to get maximum value in the shortest amount of time.