Extending Enterprise Search with Federated Connectors

Raritan Technologies has developed a suite of Federated "connectors" which enable rapid deployment of Federated Search solutions which can bring content together from the many content silos in an Enterprise by:

- Providing a "code once, configure often, customize rarely" strategy. Raritan has developed a core set of configurable workers for HTTP/HTML, SOAP, SQL and Z39.50 protocols which can be configured to hit many search sites with minimal effort. The workers are extensible when needed to handle unique situations.

- Providing a standard way of accessing secure web sites using a Secure Processing layer. This enables search sites which enforce security through authentication, login, IP validation or HTTP referrer to be included in a Federated search. The security layers handle various kinds of session management such as cookies, rewritten URLs, hidden field "chaining" in a transparent fashion.

- Providing a User Interface management layer which allows for single sign-on. Once the user is identified, the search sites that they have access to is dynamically selected so that each user can have a personalized federated search experience.

- Providing a sophisticated configuration layer which enables a more powerful "consolidated" user interface to enhance the user experience and to minimize the time required to set up a multi-field search.

Why Federated? With the explosion of information on the internet and intranet, search tools that can rapidly filter mass quantities of data to focus on relevant or timely information are highly valuable. Along with sheer volume goes the inevitable profusion of data formats and competing search services which both enhance but paradoxically inhibit searching. The enhancement is due to the fact that more tools are available; but at the same time, the wide array of choices exacerbates the problem that search software was originally designed to solve! Given many options, the first question is –"Which source has the information that I want? In other words, "I know what I'm looking for, but I don't know where it is – that's what I need a search tool for". A search system that requires too much information from the user about where and how it should look becomes self-contradictory. Federated search attempts to minimize this problem by multiplying the effect of a single search - effectively eliminating this classic "Catch 22" of search.

In a federated search, one query is distributed to many data sources. Ideally, the system hides the mechanics of searches (data source, security, query format, data format etc.) from the user but also does a lot of work to minimize the required user input. The user knows who they are (single login) and what they want to find, the system should then determine what sources are available and which sources – whether they are local files, relational databases, enterprise search collections, web sites or bulletin boards - are likely to have the relevant information. The federated system manages the query distribution, applying security information where appropriate and combines results from local and distributed sources- web sites, databases, and file systems to provide a consolidated result set. The goal is to make search easy and convenient – a one-stop search "Portal".

Raritan's Search Integration Framework Toolkit (SIFT) provides connectivity to many Enterprise Search Engines and supports:

- Query distribution and result consolidation using an extensible Java object model. The result set is based on XML, which provides many opportunities for customization and extension.
- Handles multi-threading (parallel processing) of search
- Integrates with ranking, sorting and categorization tools.
- Easy integration with many Enterprise search tools – (Microsoft SearchServer, Google OneBox, Autonomy/Verity/Ultraseek, FastSearch, Endeca, Exalead, Dieselpoint, Lucene, Omnifind)
- Configurable with XML.
- Supports J2EE standards.

There are some problem areas that must be addressed.

Federating secure access: An inherent problem with federating is that different search sources may require different security credentials. The Raritan Technologies solution includes a configurable security layer. Connectors that access secure sites are required to handle security individually. What is needed is a way to provide a single sign-on capability so that the user a) does not need to provide all of their security information every time they search and b) so that different users can have personalized access to search data. In addition, although there are a number of ways to access secure data, it is possible to develop standard solutions for the most popular security mechanisms to maximize opportunities for code reuse.

Coding new connectors for every search site is time consuming: If new code must be written for each new worker, there is a high cost in development, testing and maintenance. What is needed is a system that is adaptable, easily extensible and maintainable – this means putting as much information as possible in a "configuration layer" that is easily managed and extended to handle new situations. This separates the mechanics of generating search queries and formatting results from the details of the query language and native result format. Configuration provides the "mapping" between the native query/result formats and the common processes of the federated infrastructure.

Solutions:

Raritan technologies has developed a suite of configurable federated workers that allow the Connectors to be rapidly deployed and targeted to search sites which use HTTP, SOAP, SQL, OpenSearch and Z39.50 protocols. A configurable worker is a reusable piece of code that can "federate" within a common protocol (for example HTML) – where the request/response "language" is the same and the data retrieved has the same general format but differs in its detailed structure. For example, the exact layout of a set of database tables or an HTML page varies across databases and web pages respectively even though, the same transmission languages (SQL for databases and HTTP for web sites) and the same data formats (row, column, value and HTML) are used. The Raritan solution provides a common code base to deal with the problems common to each protocol type and a sophisticated configuration layer to manage the differences. This enables new search sites to be added to a federated solution by adding a new configuration "mapping".

The configuration layer also allows "intelligent mapping" of queries to sources so that the user does not have to know what is the best source for a particular request. This also allows Federated connectors to execute complex

coordinated searches from a single user interface by consolidating search fields with common meaning but different syntax across sites while also including site-specific fields – which cause search to be automatically focused on the sites which support them. The user concentrates on what they want to find, not where to find it – that is the software's job.

The solution also includes a customizable security layer to handle access to non-public search sites that can be "personalized" so that each user has independent access privileges to data sources.

Configurable HTTP Connector

- Configuration determines mapping from User Interface to "get" or "post" HTTP request formats.
- Pluggable security processor – handles cookies, SSL, rewritten URLS, authentication, chained hidden parameters and other forms of Web security. Drills through security layers using security information provided by a central SecurityManager.
- XML-driven HTML scraping allows the HTML result format to be "scraped" to XML in a controlled fashion. XML can then be transformed to the common format using XSLT.

SOAP

- Accesses search "Web Services" using the SOAP (Simple Object Access Protocol).
- Transport layer is typically HTTP.
- Configuration determines mapping from User Interface to SOAP methods and parameters and from SOAP results to the common federated result format.
- Pluggable SOAPREsultProcessor to handle differences in SOAP result formats.

SQL

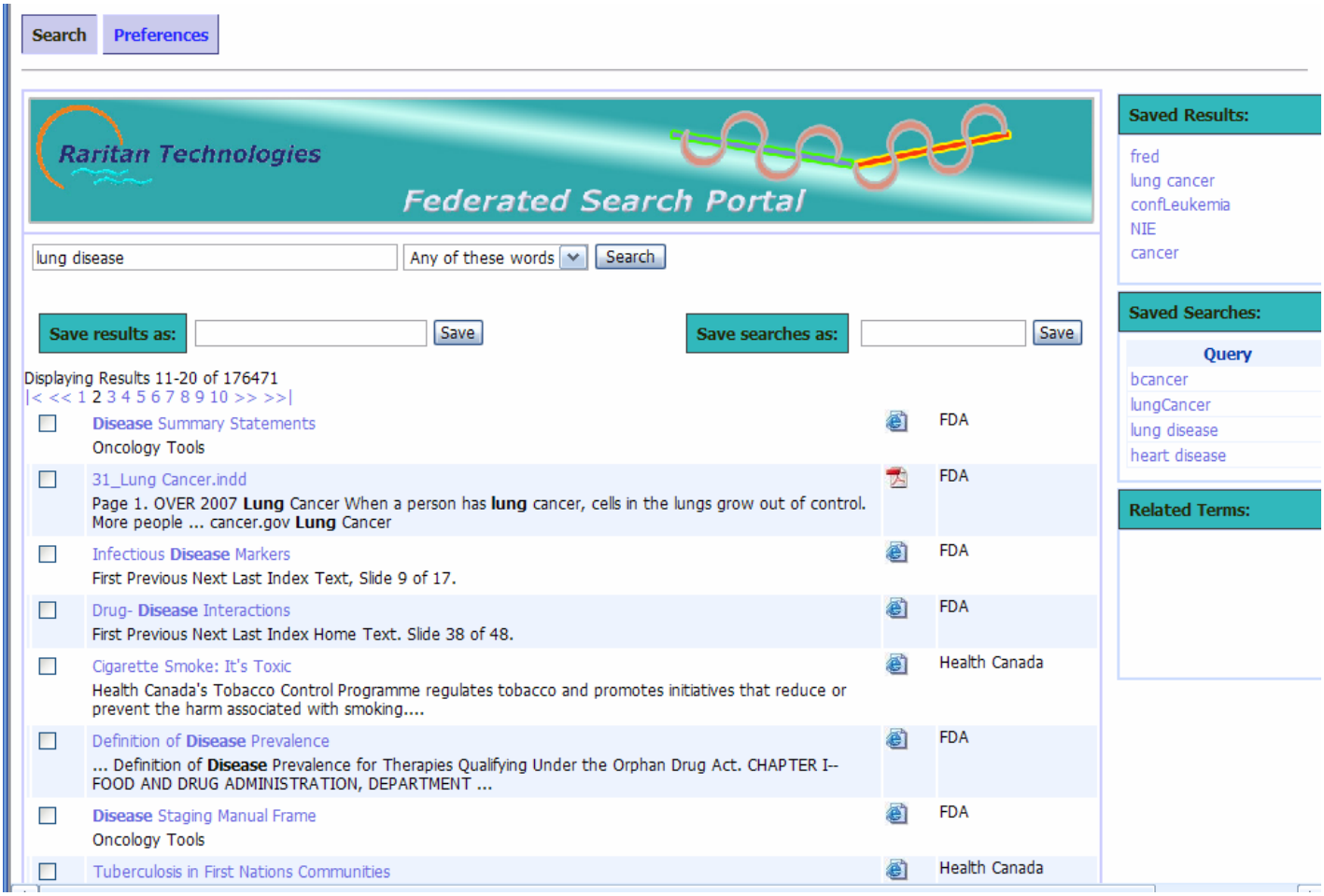
- Configurable searching of relational database systems. Configuration defines mapping of user interface fields to SQL queries and of result row/column to the common federated result format.
- Shields the user from details of database schema .
- Allows federated search of multiple databases from a single code base.

Z39.50

- Accesses sites that support the ANSI/NISO Z39.50 protocol standard.
- Provides access to a large number of corporate and university Integrated Library Systems.
- See companion white paper for more details.

OpenSearch

- This relatively new protocol is supported by an increasing number of sites.



About Raritan Technologies:

Raritan Technologies specializes in the rapid development of high-end search applications that integrate content from internal sources of information such as databases, file systems, directory services, content management systems and external sources like news services (e.g. Factiva) and web sites. Raritan's ***SIFT*** solution takes advantage of highly configurable interchangeable parts that can be targeted to the needs of a particular set of users without "reinventing the wheel" every time search is required. The result is robust, full-featured search applications.

Raritan's partnerships with Autonomy, Endeca, Iknow, Microsoft, Verity, Dieselpoint, Fast, X1 and with Factiva facilitate high-value KM solutions. The combination of content and taxonomies, Search technology and Raritan's KM and technical expertise enrich the search experience.

51 Cragwood Rd (Suite 308), South Plainfield, NJ 07080 Tel: 908.668.8181 Fax: 908.668.8870
sales@raritantechnologies.com www.raritantechnologies.com