

What's in a Name?

Federated search goes by a number of different names. *Metasearch (or meta-search), distributed search, directed search, broadcast search, deep web search, cross-database search, and universal search* are often, but not always, used synonymously with “federated search.” Metasearch is a term that is often used to refer to a search engine that searches other major search engines. Dogpile, for example, is dedicated to searching the three big search engines: Google, Yahoo!, and MSN. Some would argue that metasearch engines aren't federated search engines because, even though they search the underlying search engines in real time, the underlying search engines may not have the most current information since they themselves are “crawlers.”

Other Important Features of Federated Search

Three additional features are highly desirable, but not part of everyone's definition of a federated search. They are aggregation, ranking, and de-duplication (or “dedup'ing”), defined as follows:

Aggregation - Aggregation is the process of combining search results from the different sources in some helpful way. A federated search engine might present all of the results from one source then, beneath those results, present the results from the next source, and so on. Aggregation may incorporate sorting (e.g., by date, title, or author), or it may involve ranking, also known as relevance ranking.

Ranking - A researcher searching a couple of dozen sources via a federated search engine usually wants to know which results are most relevant to his or her search from among all of the sources. Relevance ranking compares results from all sources against one another and displays the results in order. Surprisingly, not all federated search engines rank their results. This is largely because ranking is difficult to perform well.

De-duplication - A federated search engine may retrieve the same result or document from multiple sources. Users are not interested in seeing duplicate results, yet it turns out to be difficult to remove duplicates effectively. Two documents may have the same title and author, but might actually be different revisions of one document. How does the federated search engine decide which document, or documents, to return? Like ranking, de-duping is a challenge.

Part III—Federated Search: Beyond the Basics

We have only scratched the surface of the technology. I recommend playing with federated search applications and reading some of my other articles.

Some federated search applications include:

- * [Mednar.com](http://www.mednar.com) - Searches medical information sources.
- * [Biznar.com](http://www.biznar.com) - Searches business-related sources.
- * [WorldWideScience.org](http://www.worldwidescience.org) - Searches science content from all over the world, from government agencies, as well as other quality research and academic organizations.
- * <http://search.smartlib-bibliogen.ca/zengine?VDXaction=ZSearchSimple> - Searches Capital SmartLibrary Consortium of Libraries.
- * <http://osulibrary.oregonstate.edu/metafind/about.html> - Searches Oregon State University's Library.
- * <http://scinceroll.polymeta.com/search/ui7/searchfr.jsp?un=scinceroll> - Searches a medical student's journey inside genetics and medicine through web 2.0.
- * [Science.gov](http://www.science.gov) - Searches science documents from a number of US federal government agencies.
- * <http://lifesearch.indexdata.dk/#> - Searches University of Copenhagen's Library of Faculty of Life Sciences.
- * [Scitopia.org](http://www.scitopia.org) - Searches digital libraries of leading science and technology societies.
- * <http://www.techxtra.ac.uk> - Searches 31 different collections relevant to engineering, mathematics and computing, including content from over 50 publishers and providers.



Deep Web
TECHNOLOGIES

Federated Search Primer

Written by: Sol Lederman

Part I— Federated Search Finds Content that
Google Can't Reach

Part II— A Definition of Federated Search

Part III— Federated Search: Beyond the Basics

301 N. Guadalupe, Suite 201
Santa Fe, NM 87501
505-820-0301
505-983-7621 fax
deepwebtech.com
info@deepwebtech.com

Part I- Federated Search Finds Content that Google Can't Reach

Federated search facilitates research by helping users find high-quality documents in more specialized or remote corners of the Internet. Federated search applications excel at finding scientific, technical, and legal documents whether they live in free public sites or in subscription sites. This makes federated search a vital technology for students and professional researchers. For this reason, many libraries and corporate research departments provide federated search applications to their students and staff.

To really understand what federated is and how it works we should first provide some background.

Crawling the Web: How Typical Web Search Engines Work

There are two basic approaches to finding content on the Web. The approach that Google and all major search engines use is to “crawl” the Web. Google, over many years, has amassed a list of billions of Web sites. In the early days, it's likely that Google learned about many Web sites when owners registered their sites with them. Today, Google can find new Web sites through links from sites it already knows about. Google periodically visits the sites (and the sites' pages) on its list and identifies the links at that site. It then follows each link it finds to arrive at other pages where it starts the process over to find more links. In doing this, Google discovers sites it didn't know about during previous visits. This process of going from one page to another and then to another is referred to as “crawling,” just like a spider crawls from one thread to another in its web. In fact, Web “spiders” are commonly referred to as “Web crawlers.” When you create a new site, just create a link to it from another site, or get someone to do it for you, and Google's crawler will discover it.

The trouble with crawling is that this search technique doesn't find everything. One might believe that through sufficient crawling, one could find all Web pages. In fact, only a small percentage of the Web's content is accessible to Google. The term “deep Web” refers to the vast portion of the Web that is beyond the reach of the typical “surface Web” crawlers. Surface Web search engines like Google can't easily fathom the deep Web because most deep Web content has no links to it. How can that be? Consider this example: Let's say that you are researching the effects of some chemical or hazardous substance on humans. You would be well advised to search the National Library of Medicine's [Toxicology Data Network](#). Most of the information you would find there you would not find via Google. Why? Because, to find the research articles, you would have typed one or more words in a search box and you clicked on the “search” button. Few, if any, of the articles you found had links to them from any Web site. Google couldn't find those articles because Google isn't designed to fill out search forms and click “submit” the way humans do. In particular, Google wouldn't know what search words to put into the form. Additionally, even if Google did know what to enter into search forms and how to submit them, Google wouldn't be able to retrieve all of the documents from the source. This would leave Google with very incomplete content from deep Web sources.

What Makes Federated Searches Different? It's About the Search Forms

While in most cases, Google doesn't fill out search forms, this is exactly what federated search applications (also known as federated search engines) do. Why doesn't Google fill out forms? It turns out that filling out forms is a difficult problem. Federated search engine builders have to customize their search software for each Web form they encounter. While Google has a general approach to crawling links from any Web site, federated search engines are programmed with intimate knowledge of each search form. The specialized software must know not only how to fill out the form and how to simulate the pressing of the “search” button, but also how to read the results that the Toxicology Data Network (as in the example above), or any other source, provides. Both are difficult to do well.

The benefits of Federated Search

The essential benefits of federated search to its users include efficiency, quality of search results, and current, relevant content.

Efficiency, Time Savings

Using a federated search engine can be a huge time saver for researchers. Instead of needing to search many sources, one at a time, the federated search engine performs the many searches on the user's behalf. While federated search engines specialize in finding content that requires form submissions to retrieve, it isn't the only criterion for being a federated search engine. A federated search engine also associates content from different sources. Federated search uses just one search form to cover numerous sources, and combines the results into a single results page.

Quality of Results

Federated search engines show their value best in environments in which the quality of results matters, such as libraries, corporate research environments, and the federal government. In the case of the federal government, the constituents of the government benefit greatly from such applications. A major difference between a federated search engine and a standard search engine like Google is that the client who contracts for the federated search service selects the sources to search. In almost every case, the sources will be authoritative. Google, on the other hand, has very minimal criteria for source selection. If a Web page doesn't look like outright junk (i.e., spam) Google will present it among the search results. Thus, the federated search engine acts as a helpful librarian does, directing users to excellent quality.

Most Current Content

In addition to filling out forms and combining documents from multiple sources, another important benefit of federated search engines is that they search content in real time. Real time data is crucial for researchers who are searching for up-to-the-minute content or for content that changes frequently. As soon as the content owner updates their source, the information is available to the searcher on the very next query. By contrast, with standard search engines/Google, the results are only as current as the last time that Google crawled sites with content that matches your search words. Content you find via Google might be days or weeks old, which can be fine depending on your situation, but can be problematic if you want the most current information.

Part II—A Definition of Federated Search

While not everyone agrees on all details of what federated search is, here's a fairly well-accepted definition:

Federated search is the process of performing a simultaneous real-time search of multiple diverse and distributed sources from a single search page, with the federated search engine acting as intermediary.

This definition is a mouthful, isn't it? Let's look at the key words in the definition and their influence on the value of federated search:

- * **federated** - Content is combined from different sources saving the effort of searching sources one at a time.
- * **simultaneous** - Federated search queries all user-selected sources at once. It would be unacceptably slow if it waited for all of the results from one source before querying the next.
- * **real-time** - Federated search occurs live and results are current. There's no stale content.
- * **multiple** - The value of federated search to the researcher increases as the number of sources increases.
- * **diverse sources** - Federated search engines typically can search sources containing documents of different types, e.g. PDF, Word, PowerPoint. The process of extracting text from documents of different types is hidden from the user.
- * **distributed sources** - Federated search engines expect to search content that lives in different locations.
- * **single search page** - Federated search engines provide a single point of searching.
- * **federated search engine acting as intermediary** - The federated search paradigm is such that the user doesn't communicate directly with the content sources when performing searches. The user submits a search to the federated search engine which, in turn, submits the search to each of the content sources. Each content source provides its results to the federated search engine which combines all of the results from all the sources into a single page of results. Note that federated search was developed independently of the Web, and therefore federated search engines need not be Web-based.