Science Research: Journey to Ten Thousand Sources

Abe Lederman
President
Deep Web Technologies
Santa Fe, New Mexico
Member, SLA

Abstract

The future of scientific research depends on sifting through more information, more quickly, and more effectively. For a researcher to expect to be able to search 1,000 databases simultaneously for critical information is not unreasonable. While parallel search is the domain of federated search, the current paradigm has severe limitations. The limitations, which include speed, relevance ranking, and selecting the appropriate sources, become painfully obvious when one attempts to search more than a few dozen sources simultaneously. A paradigm is needed for scalability to not only overcome the limitations but also help us assimilate important information.

Scalability matters because, for the first time in history, we have access to diverse research from many of the world's nations and we are too overwhelmed to make sense of it all. Scalability also matters because the acceleration of the scientific discoveries we need to make to solve the world's pressing problems depends on the cross-fertilization of ideas that occurs when scientists from different disciplines contribute to, and draw from, a large number of information sources.

How is massive scalability achieved? Deep Web Technologies, in partnership with the U.S. Department of Energy's Office of Scientific and Technical Information, has developed a divide-and-conquer approach. While a 1,000-source search engine may struggle with scalability issues, combining 10 to 20 federated search engines that each search 50 to 100 sources allows each search engine to perform a manageable amount of work.

This paper discusses scalable architecture. Our experience with its implementation in several science research portals, such as WorldWideScience.org and ScienceResearch.com, is described. Finally, our vision of what will be needed to create a future in which searching 10,000 information sources in parallel is achievable is discussed.

Introduction

In February 2006, Dr. Walter Warnick, Director of the U.S. Department of Energy Office of Scientific and Technical Information (OSTI), introduced the concept of global discovery to the attendees of the AAAS Annual Meeting (Global Discovery Introduced at AAAS 2006). He made three important observations:

1. Scientific progress depends on the diffusion of knowledge;
2. Knowledge that may lead to breakthroughs frequently resides in distant scientific communities; and
3. Innovation is needed to speed up the diffusion of knowledge.

Dr. Warnick shared his vision of creating a global discovery facility capable of supporting the diffusion of knowledge, which would, in turn, lead to an acceleration of scientific discovery. (Global Discovery: Increasing the Pace of Knowledge Diffusion to Increase the Pace of Science).

> *Our ultimate goal is to have a true Global Discovery facility. To help create it, we have undertaken a number of activities which, collectively, we call Innovations in Scientific Knowledge and Advancement, or ISKA. The Global Discovery facility would aggregate, search and rank all of the important, Web-accessible databases. It has the same goal as the fabled Library of Alexandria, namely to make all of science available in one place. Except in this case the place is everywhere at once, because anyone in the world could access the Global Discovery facility.*

Eighteen months later, in June 2007, Dr. Warnick realized a significant achievement toward his vision: The U.S. Department of Energy and the British Library, together with eight other nations, launched WorldWideScience.org, a global gateway for science. (Global Science Gateway Now Open).

WorldWideScience.org ushered in a new era in scholarly research. The launch of the global gateway made possible the simultaneous search of scientific and technical information from multiple countries from a single search box. Today, 55 countries have contributed a total of roughly 375 million pages of science information to WorldWideScience.org. The participating countries represent approximately 73% of the world's population and demonstrate the importance of global collaboration in disseminating information to help advance science (China Joins the WorldWideScience Alliance).

The global collaboration making up WorldWideScience.org helps to advance science by exposing researchers to developments beyond their borders. The global science community is

growing and not just in the western world. China, for example, is second only to the United States in the number of research papers published (Is the United States Losing Ground in Science?). Other Asian countries are also increasing their prominence. International science is critical to monitor because such science represents a diversity of perspectives and approaches that complements the ways of thinking of any single nation. By relying on only the content available from the major publishers and aggregators, researchers miss other important content, in particular the output of scientists who do not publish in mainstream journals. The world is shrinking, the brain pool is growing, and the output of science is everywhere.

Deep Web Technologies, founded by Abe Lederman, also presented at the 2006 AAAS Annual Meeting. The presentation discussed how significant elements of Dr. Warnick's vision could be implemented (Global Discovery: Turning Vision into Reality). Deep Web Technologies was well qualified to carry forward with the global discovery work because we had already developed sufficient scalable search technology to implement the search functionality for Science.gov, the portal to U.S. government scientific and engineering information.

Beyond Science.gov, Deep Web Technologies built the search technology for WorldWideScience.org utilizing a more modern approach to scalability, which is discussed in this paper. We have also made substantial progress in launching our own portal, ScienceResearch.com. This portal aims to unify the World Wide Web's dispersed science to become the world's most comprehensive portal for science. Additionally, the portal seeks to make "long tail science," the very specialized science that may appear to be of limited interest, available to a larger audience through which applications may be found. Hopefully, the portal will serve as a catalyst for scientific discoveries and innovative solutions to many of the world's pressing problems.

In order to build a massive portal for science, we must address a number of challenges; some are technical while others are administrative or political. We have solved some of the problems yet we are still in the first steps of our journey to searching 10,000 sources simultaneously. This paper discusses the problems we aim to solve, our approaches to solving them, where we are today, and what we envision the future to look like. In our discussions of the challenges and approaches to scalability, we consider the near-term goal of searching 1,000 sources, which we expect to achieve within one year. In sharing our vision, we consider what will be required to meet our long-term goal of managing 10,000 sources in a single search.

Challenges: Overview

The volume of scientific information – the output of the global R&D performed by government laboratories, educational institutions, and corporations – is huge and growing. Not

only is the volume of content growing, but the content is becoming more dispersed, given the decentralized model for document storage on the Internet. While much of the desirable scientific information is organized in repositories, the repositories are scattered about and no one knows where all the repositories are. The repositories to which American scientists are commonly exposed are those of the large publishers (e.g., Elsevier and Springer) and those offered by major aggregators (e.g., Scopus, Ingenta, and Web of Science). While historically, major publishers and aggregators may have delivered most of the noteworthy science, the open science movement and the proliferation of open access journals has led to the creation of many other channels to credible science information. MIT, for example, recently announced that it will "freely and publicly distribute research articles they write." Stanford's School of Education and several departments at Harvard University are following suit. (MIT Will Publish All Faculty Articles Free In Online Repository). The value of these new channels is especially immeasurable to researchers in poor countries who cannot afford to subscribe to content from the major publishers.

In considering ways to find and provide access to credible science, we first consider Google and the other major search engines. While Google and others facilitate finding disparate information on the Web, the majority of Web content is actually beyond their reach, in the hard to access region of the Internet known as the Deep Web. In particular, the Deep Web contains large amounts of high-quality scientific, technical, and business information, which is not accessible to the popular search engines. Thus, Google and others can provide only limited access to the full range of scientific information. Harvesting and aggregating disparate Deep Web content is also not feasible for the majority of Deep Web sources since most content sources do not provide a mechanism for harvesting their content. This situation makes federated search (including federation of Google content) the best option and, in many cases, the only option.

Federated search works very differently than does the web crawling (Google) approach. Crawling works by following links from one web page to another and indexing the contents of each page found. While Google has amassed an index of billions of web pages, Google cannot usually see content inside databases, which are only accessible by filling out search forms on web pages. Federated search applications utilize custom-built pieces of software called "connectors" that know how to interact with the search format of Deep Web databases to submit queries, retrieve results, and extract the relevant fields from the results (e.g., title, author, and abstract). Federated search applications utilize different connectors with different databases. Federated search applications also commonly aggregate results from the different searched databases and sort the aggregated results in order of relevance.

Figure 1 illustrates the flow of execution of the major steps in a federated search. A user enters a query. The query is submitted to multiple search engines in parallel, results are retrieved

from each of the sources, the results are aggregated and relevance ranked, and then, finally, the results are displayed to the user.

Several characteristics make federated search particularly suitable for performing research:

- Federated search accesses Deep Web databases that Google cannot crawl and index;
- Federated search queries multiple databases in real-time providing users the most current content available; and
- Federated search usually accesses only high-quality databases, freeing users from having to filter out results from poor sources.

Federated search technology was developed specifically to mine Deep Web content. While locating content web crawling cannot, federated search does have some limitations, particularly related to scaling the number of content sources searched as well as supporting a large number of simultaneous searches.
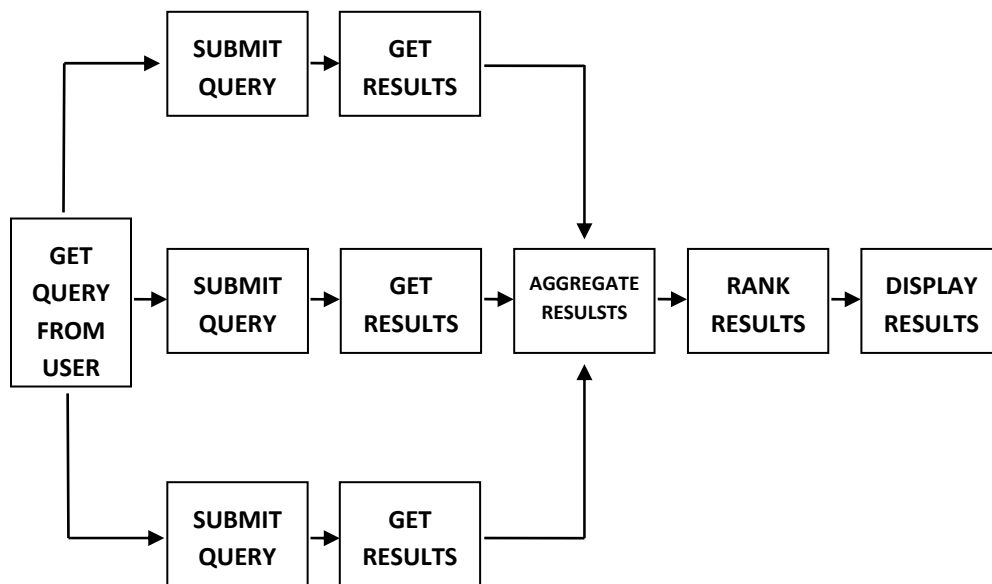

Figure 1: Federated search flow of execution.

Scientists build on the work of their peers and predecessors. Additionally, scientists benefit from the cross-fertilization effect of utilizing the work of those in different fields. Key to building on the research of others is the ability to find useful results. Current federated search applications limit researchers by not being comprehensive, forcing those who want to be thorough in their research to search multiple applications one by one, whether they are federated search applications or not. Because of the inconvenience of searching different repositories and because researchers may only know about the ones their institutions subscribe to, researchers are missing good content. How do we deal with the fact that, for a sufficiently broad discipline, there

may be hundreds, potentially thousands, of content sources providing relevant information if related fields contributing to cross-fertilization are included?

The obvious solution to the problem of organizing search for dispersed Deep Web content is to build federated search engines capable of searching a larger number of sources. The problem with this approach is that traditional federated search engines were not designed to scale. The current paradigm that works well for searching four or 40 sources breaks down as the number of sources approaches 100. When we attempt to scale federated search to 1,000 sources, new problems are encountered.

Figure 2 identifies a number of the major challenges we must address in the context of the flow of execution of scalable federated search (see, for example, the Challenges: In-Depth Discussion section). Figure 2 shows that user requests are examined to determine if cached results can satisfy the search. If so, the search need not be performed. Figure 2 is simplified by not considering search results associated with specific sources so the cache may retrieve results from a subset of the selected sources. If cached results are not available, then the federated search engine selects appropriate sources based on the query terms, submits the query to multiple sources in parallel, applying load-balancing techniques to throttle traffic to individual sources, retrieves results, aggregates and ranks the results for relevance, and finally, after filtering, displays the results.
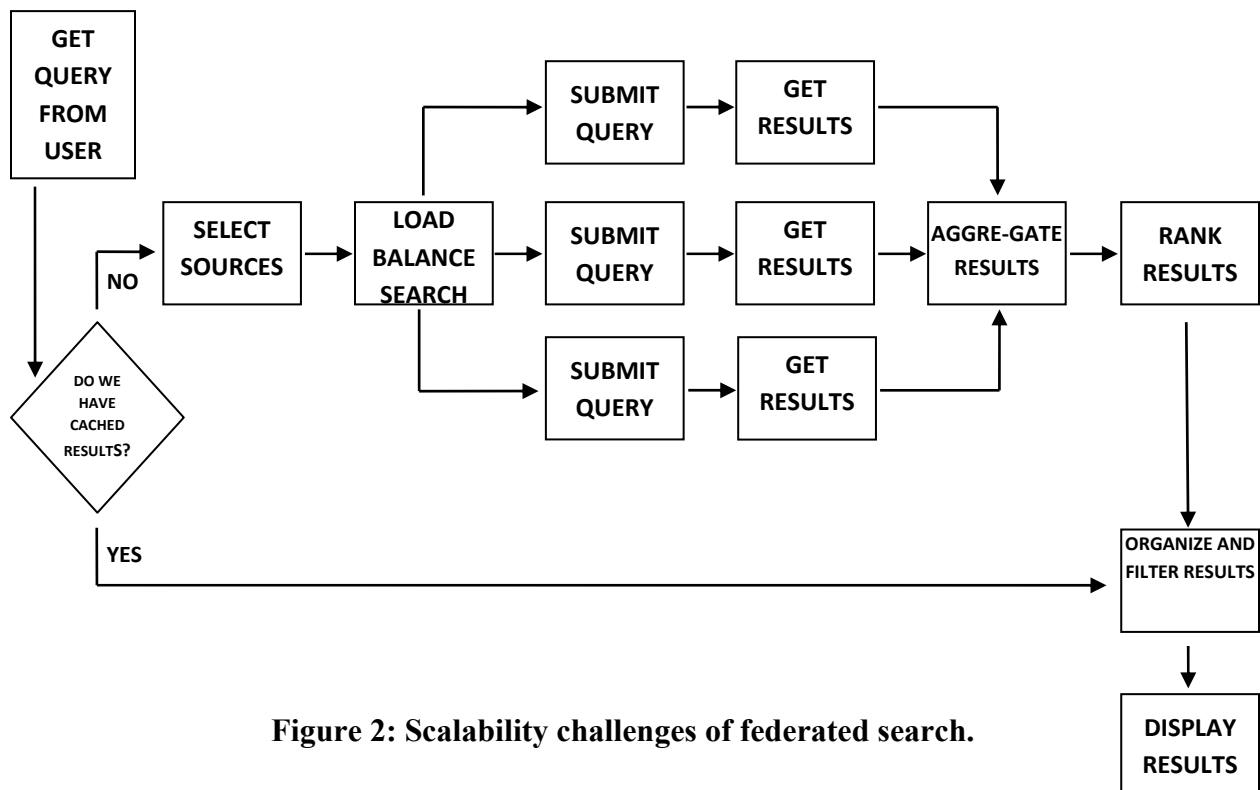
Figure 2: Scalability challenges of federated search.

Challenges: In-Depth Discussion

We have identified five major challenges to scalable search.

How do we facilitate source selection? In a 1,000-source federated search application, searches are performed in real-time, with each source selected for a user query resulting in substantial traffic to the source. Minimizing the number of searches to sources not likely to return relevant results is important. We are experimenting with developing source-selection technology to select sources for users. Sources selected will be based on historical query and result information. If we have evidence that a previous user who performed the identical query got relevant results from particular sources, those sources will be selected. Our approach also includes the use of a thesaurus to identify historical queries different than, but synonymous to, the query we are examining. Matching user queries to the best sources not only lessens the burden on the sources, but guides users to relevant sources, especially ones they may not have considered. Source-selection technology will be particularly important to implement when the number of sources is so large that unassisted source selection by the user would become an overwhelming task. We want to avoid situations in which users, in frustration, select all sources or give up. Of course, we would also group sources into categories to facilitate selection and discovery, although a major benefit of source-selection technology is the selection of sources in categories the user might not have considered. For advanced users, a mechanism to override the sources chosen by the source-selection system will be provided. We will not, however, let any user select all sources. The number of sources selectable for a query will be limited.

How can we ensure that results are relevant and well organized? As federated search applications scale, more results will need to be processed. Thus, the application must manage these results to avoid overwhelming the user. Relevance ranking will need to work well in any scalable search environment (see, for example, the Divide-and-Conquer Approach Supports Scalability section). Additionally, federated search applications will need to have powerful sorting, filtering, query refinement, and visualization features, such as clustering, to facilitate navigation of the results by the user.

How can the federated search engine manage traffic to sources? We must consider the likelihood that as federated search applications scale to search more sources, the applications will prove more useful to researchers and will attract more users. We must also consider that source-selection technology, while important, may not prove sufficiently effective in minimizing traffic to sources. We are exploring the feasibility and value of three approaches to reducing the number of search requests made to sources.

1. **Caching of queries and associated results**. Caching involves storing the results from popular queries. When a user performs a search on one of those queries, results from the cache can be provided, thus decreasing the burden on the source. There are two considerations to employing a cache. First, the queries that are submitted with sufficient frequency to benefit from caching need to be identified. Second, a mechanism to refresh the cache when its contents no longer reflect the results provided by each source needs to be developed. Developing a cache refreshing mechanism creates the need to know when cache contents become out of date. We expect the implementation of caching to be a straightforward task. We will build a least-recently-used (LRU) cache model to remove old, infrequently referenced search results from the cache when the cache is full, thereby making room for new search results. By using a cloud-computing environment, we can experiment with varying amounts of storage for the cache.

2. **Providing canned queries where appropriate**. Users may be helped by the federated search application recommending queries on different subjects. Creating canned queries makes sense for the popular searches for which we have cached results. This approach, as well as the previous one, works well if there are sufficiently many popular results. Results of canned queries can also be made available to scientifically oriented websites to complement their existing content.

3. **Controlling the flow of queries to individual sources.** The most complex approach to minimizing traffic to individual sources, but the one which may have the greatest payoff, is to throttle search requests so queries are never submitted faster than some given rate for each source. Implementing this mechanism requires monitoring each source for a degraded response to queries, which may vary at different times of the day and during the week. The application would throttle back on queries during peak times, which may be different for particular sources. The application would also monitor the effect of its throttling and adjust future throttling accordingly. An interesting variation on the throttling of search requests to sources is to perform canned query searches at off-peak times for different sources and to cache the results.

How does the federated search engine handle the computational load? This problem is perhaps the easiest of the five major challenges to manage. As federated search applications scale, more search and more search results will need to be processed. More aggregation of results and more relevance ranking will also need to be preformed. While one approach is to distribute the workload across machines, a simpler approach is to increase the capacity of the federated search server by adding processors, storage, or memory or by upgrading hardware. Another way to overcome hardware limitations is to perform federated search in a cloud-computing environment, in which resources can be dynamically allocated to match the application loads.

How do we find, build, and maintain connectors for 1,000 sources? We have developed processes and software to build, test, monitor, and repair connectors. As the pool of connectors gets larger, our existing tools will need to be refined to assist those who monitor the connector status reports by providing diagnostic information. While today's monitoring tools alert connector administrators when problems are encountered or when source search interfaces change, tomorrow's tools should perform tests of problematic sources and, ideally, provide sufficient information to facilitate rapid updating of connectors. We are exploring ways to partially automate the creation of connectors. These approaches, if they prove fruitful, will also assist in the connector maintenance process. The challenge of finding quality connectors is a major one, which is discussed in the A Vision for the Future section.

Divide-and-Conquer Approach Supports Scalability

We have developed and put into practice an approach to scalability that divides the computational and network aspects of search, retrieve, and rank into tasks distributed among multiple federated search engines. Figure 3 illustrates the multi-level approach we use to break up a search of many sources into manageable smaller searches. For example, Server A initiates a search of 200 sources. Server A asks both Server B and Server C to each search 100 sources. Servers B and C in turn ask Servers D, E, F, and G to each search 50 sources. As servers D, E, F, and G obtain results from their designated sources, relevance ranking of the received results is performed and the best results are passed to Servers B and C, which in turn roll up their results for presentation to Server A, which performs the final round of relevance ranking and presents the results to the user.
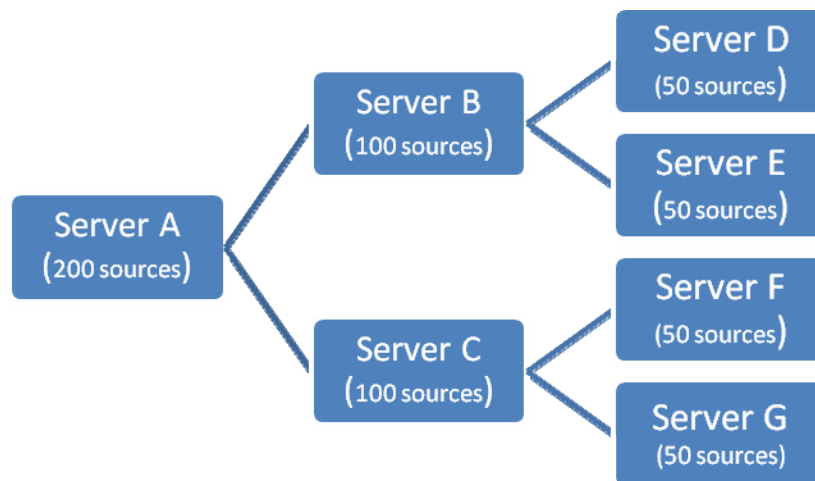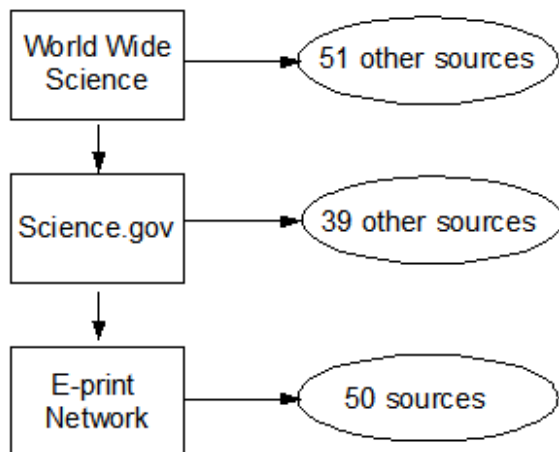

Figure 3: Multi-level divide-and-conquer approach.

The divide-and-conquer approach is transparent to users. In spreading the workload, the master server (i.e., Server A) does not ask any of the subordinate federated search engines to

perform more computations than they are capable of performing efficiently. This approach is very quick to implement when the top-level server (i.e., Server A) is searching federated search engines that already exist. An effective strategy we employ for building large federated search engines is to identify existing federated search engines that focus on related subjects and combine them. WorldWideScience.org is an excellent example of this approach in action.

WorldWideScience.org Models Scalability

WorldWideScience.org is a scalable federated search application. At present, this science gateway searches 52 sources. One of these sources is Science.gov, a federated search application. The other 51 sources are individual databases. Science.gov searches the E-print Network, a federated search application, and 39 individual databases. The E-print Network searches 50 databases. This division of labor for WorldWideScience.org allows a user to search 140 sources (i.e., 51+39+50) from a single search page. Figure 4 illustrates how the search load is divided.

Figure 4: WorldWideScience.org searches 140 sources via a divide-and-conquer approach.

The Roots of Scalability

Our scalable approach was first documented in 2001 in a paper in *D-Lib Magazine*, which was jointly written by Abe Lederman and a number of OSTI staff, including Dr. Warnick (Searching the Web: Directed Query Engine Applications at the Department of Energy). The divide-and-conquer approach was described in reference to the federated search technology of Innovative Web Applications (IWA), a predecessor company of Deep Web Technologies.

"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."

The cascading hierarchy was not exploited until later.

Implementation has its roots in work we performed for the U.S. Department of Energy under a Small Business Innovation Research (SBIR) grant starting in 2003. A next generation federated search engine, Explorit, was built as a set of a dozen Web Services, which allows distribution of computation, network, and storage loads in much the same way as cloud computing does today. We proved the feasibility of a federated search architecture that allowed for partitioning and distributing the various tasks involved in performing a federated search. By late 2005, this scalable, distributed architecture was deployed in Science.gov. In 2006, our scalable approach to OSTI's Science Accelerator portal was introduced. In late 2006, working closely with the Intel Corporate Library, the largest deployment of Explorit at the time was successfully launched to Intel's 100,000 employees worldwide. Intel's experience deploying federated search is documented in a case study presented at the 2007 SLA annual Conference and published in Information Outlook (Hill 2007). In 2007, WorldWideScience.org was launched, which federates content from Science.gov, which in turn federates content from the E-Print Network.

ScienceResearch.com Advances Scalable Federated Search

In May 2009, we launched the initial version of ScienceResearch.com, a major milestone in our journey to build the most comprehensive science portal in the world. As of this initial launch, ScienceResearch.com directly searches about 190 unique science sources, and through the divide-and-conquer approach, ScienceResearch.com searches an additional 210 unique science sources. To the best of our knowledge, no other federated search application simultaneously searches 400 sources. In addition to the WorldWideScience.org, Science.gov, and E-Print Network portals, ScienceResearch.com also searches ScienceConferences, a portal providing access to some of the best conference proceedings, and Mednar.com, a portal providing access to literature aimed at medical professionals. Each of these portals returns their best 200 search results to ScienceResearch.com. These results are aggregated with the results returned by individual sources.

Figure 5 shows the home page of ScienceResearch.com, which provides a basic search form, much like what users are accustomed to seeing in Google. Below the search box is a set of buttons to allow a user to limit the search to Applied Science, Computation Science, Life

Science, or Physical Science content. Sources searched for each of these major areas of science include sources appropriate for the science area selected, including multidisciplinary databases and portals, such as Science.gov and WorldWideScience.org.
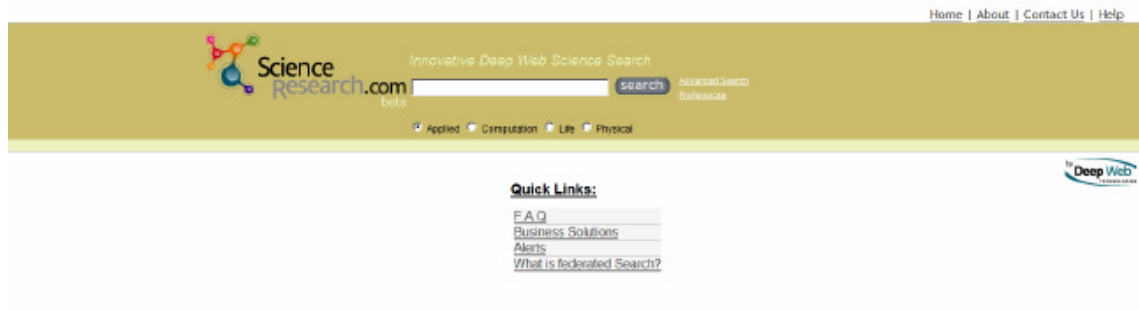
Figure 5: ScienceResearch.com home page.

Figure 6 shows the advanced search page. On the top of this search page are a number of featured sources, selected because of the scope and quality of the scientific content available through these sources. Also available below the featured sources are searchable collections of Patents and Science News. Below the featured sources and on the left of the advanced search page, searches based on specific titles, authors, and date ranges can be conducted.

On the right half of the advanced search page, a user can choose to search multiple categories of sources from a list of 13 categories, which include agricultural sciences, defense technologies, earth and environmental sciences, mathematics, multidisciplinary sources, and eight others. If a user wants to limit their search to specific sources in one category, there is a link from the category to a category-specific search page. A new capability, called SearchBuilder, will allow a user to create and save a custom search page consisting of just the sources the user wants to search.
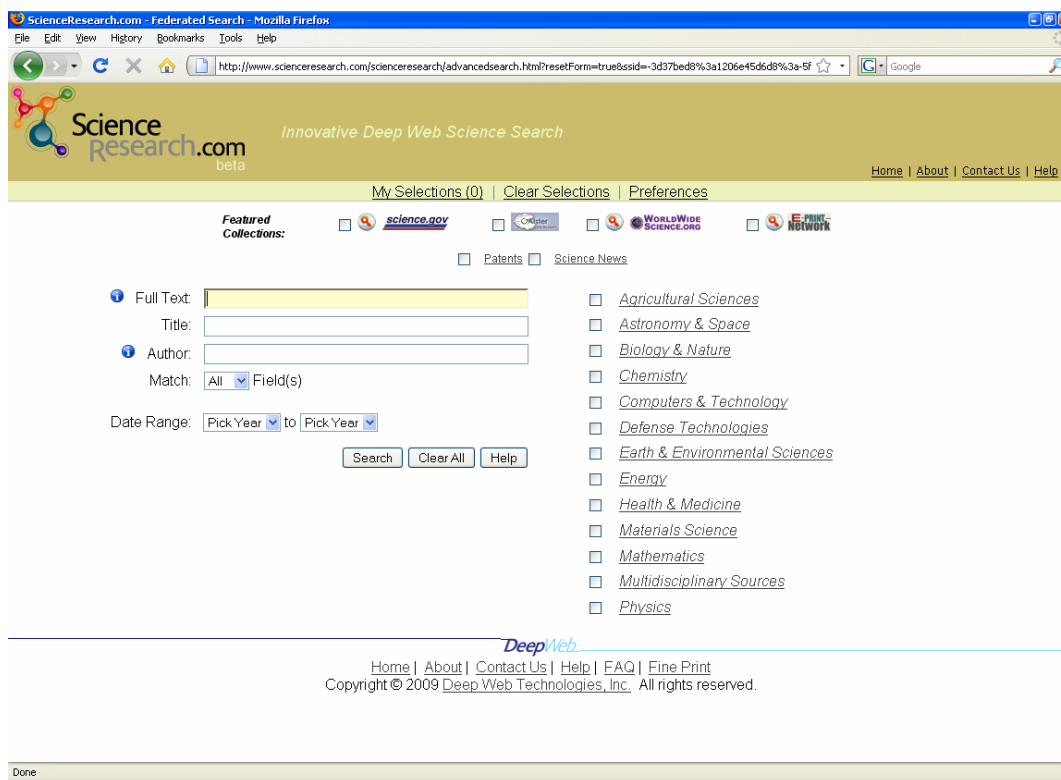
Figure 6: ScienceResearch.com advanced search page.

The search page for the Chemistry category is shown in Figure 7. Within this category, we have taken a first step toward reaching out to the scientific community to help select the best sources to search. Specifically, Grace Baysinger, Head of the Chemistry Library at Stanford University, has volunteered to serve as the Editor for the Chemistry category of ScienceResearch.com and to help select the best sources to search.

Figure 8 shows a sample results page. For this example, we searched for "green energy" in the sources in Applied Sciences category. We searched 216 sources and brought back the 1,200+ best results across all the sources searched. The results page shows the rich set of functionality provided by ScienceResearch.com, which includes a clustering capability allowing a user to browse through the set of results organized into groups of similar results. Also provided on the results page is the ability to sort results by title, author and date, as well as by rank. All retrieved results, or a selected set of results, can be e-mailed to the user or a colleague or exported to a citation manager, such as RefWorks.
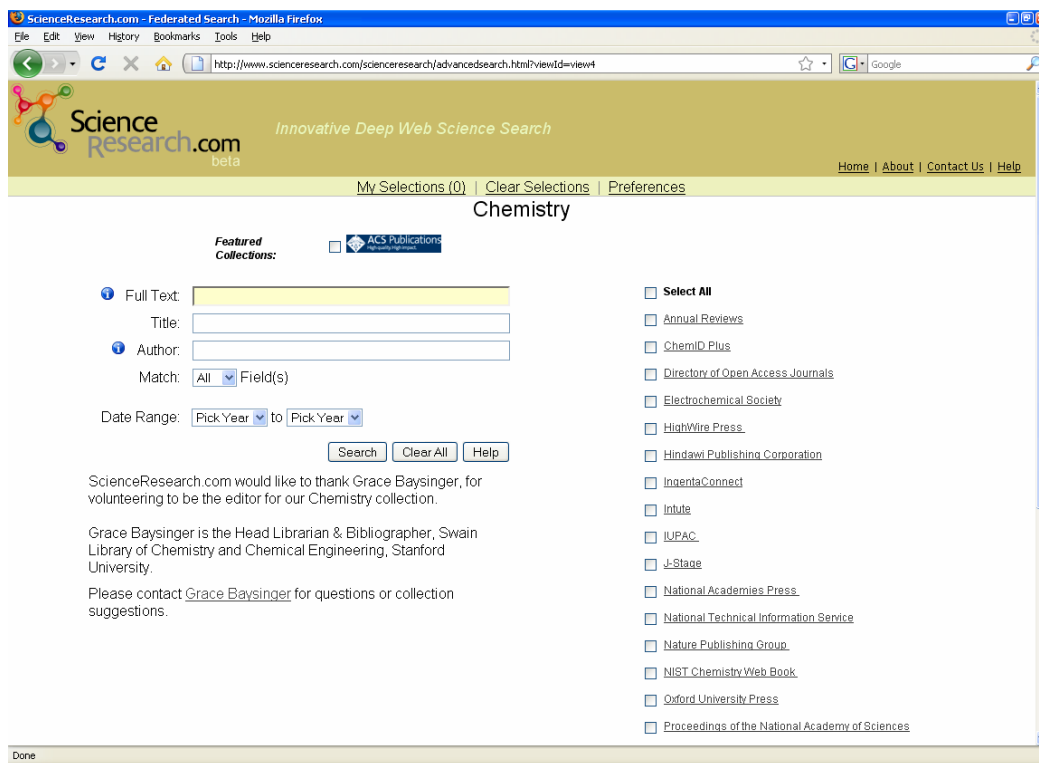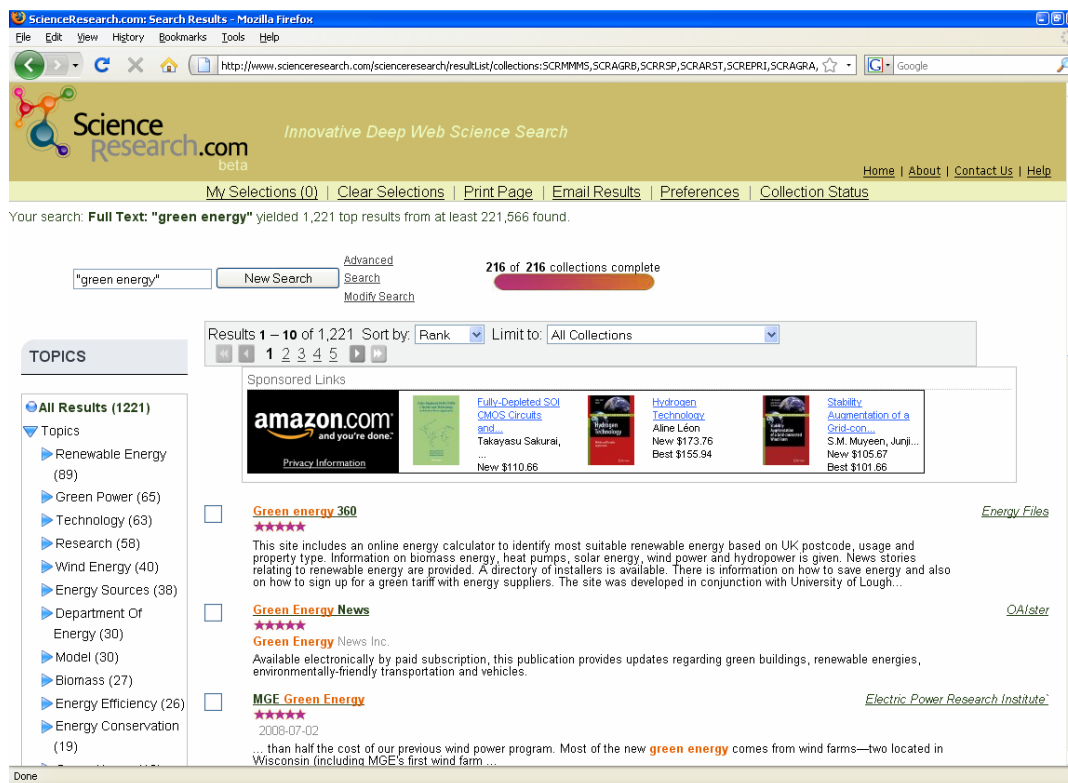
Figure 7: ScienceResearch.com Chemistry search page.

Figure 8: ScienceResearch.com sample results page.

A Vision for the Future

We have demonstrated the feasibility of the divide-and-conquer approach to scaling federated search. To reach our near-term (i.e., one-year) goal of searching 1,000 sources simultaneously, we need only to identify more relevant sources and to build suitable connectors. New sources can be combined into one or more federated search applications and can be searched by the ScienceResearch.com portal utilizing the divide-and-conquer approach.

Reaching our longer-term goal of simultaneously searching 10,000 sources is not principally a technical problem, although source selection and other technical advances considered in the Challenges: In-Depth Discussion section will become more critical.

Greatly scaling the number of sources will require the support of the international science community in several ways. The first role the community can play is to assist in the identification of sources. No single individual or organization will be able to keep up with the introduction of new sources; thus, those who have the most to gain from a comprehensive science portal, the researchers, will benefit by identifying new resources for inclusion. A second way the science community can support creation of a comprehensive science portal is to assess sources already included in the portal as well as candidate sources. In order to facilitate the rating of, and commenting on sources, a social network framework is being planned. A third value that the science community can provide is to create connectors for new sources. To facilitate this effort, easy-to-use tools and documentation will need to be developed. The incentive to the community to participate in identifying, assessing, and configuring sources may well be the creation of a portal directly supporting their individual research efforts and a feeling of being a crucial part of achieving of the overarching objective of advancing science.

While a large focus of the portal effort is on creating access to non-commercial content, the value provided by commercial publishers should not be ignored. Thus, we are committed to working out the business aspects of relationships with publishers and plan to integrate access to their content, with some of that content most likely being fee-based.

A natural path to building a 10,000 source portal is to build focused portals in numerous scientific disciplines. If, with the help of the science community, 100 subject areas each with 100 search sites can be identified, then we can use our divide-and-conquer approach to achieve scalability and combine the subject portals to form a large single portal.

Key to all of our efforts is that to accelerate science by unifying access to large numbers of content sources will require a concerted grassroots effort from a large number of individuals

and organizations who recognize the value of global cooperation and who see their role in contributing to this end.

End Notes

"China Joins the WorldWideScience Alliance: Why This Is Important," OSTI blog, http://www.osti.gov/ostiblog/home/entry/china_joins_the_worldwidescience_alliance, Accessed 28 April 2009.

"Global Discovery: Increasing the Pace of Knowledge Diffusion to Increase the Pace of Science," U.S. Department of Energy Office of Scientific and Technical Information, http://www.osti.gov/speeches/fy2006/aaas/, Accessed 28 April 2009.

"Global Discovery Introduced at AAAS 2006," U.S. Department of Energy Office of Scientific and Technical Information, http://www.osti.gov/news/2006/feb/globaldiscovery, Accessed 28 April 2009.

"Global Discovery: Turning Vision into Reality," Deep Web Technologies, http://deepwebtech.com/talks/AAAS.ppt, Accessed 28 April 2009.

"Global Science Gateway Now Open," U.S. Department of Energy, http://www.energy.gov/news/5153.htm, Accessed 28 April 2009.

Barclay Hill, "Federated Search at the Intel Library," Information Outlook, September 2007, Vol. 11, No. 9.

"Is the United States Losing Ground in Science? A Global Perspective on the World Science System," Scientometrics (forthcoming), http://users.fmg.uva.nl/lleydesdorff/us_science/, Accessed 28 April 2009.

"MIT Will Publish All Faculty Articles Free In Online Repository," The Tech, http://tech.mit.edu/V129/N14/open_access.html, Accessed 28 April 2009.

Walter Warnick, Abe Lederman, et al, "Searching the Web: Directed Query Engine Applications at the Department of Energy," D-Lib Magazine, January 2001, Vol. 7, No. 1, http://www.zbp.univie.ac.at/gj/vortrag/warnick_searching_deep_web.pdf, Accessed 28 April 2009.