Federated search as a transformational technology enabling knowledge discovery: the role of WorldWideScience.org

ABSTRACT

Purpose - To describe the work of the Office of Scientific and Technical Information (OSTI) in the U.S. Department of Energy Office of Science and OSTI's development of the powerful search engine, WorldWideScience.org. With tools such as Science.gov and WorldWideScience.org, the patron gains access to multiple, geographically dispersed deep web databases and can search all of the constituent sources with a single query.

Approach – Historical and descriptive

Findings – That WorldWideScience.org fills a unique niche in discovering scientific material in an information landscape that includes search engines such as Google and Google Scholar.

Value – One of the few articles to describe in depth the important work being done by the U.S. Office of Scientific and Technical Information in the field of search and discovery

Paper type – Review

Keywords – OSTI, WorldWideScience, ScienceAccelerator.gov, Science.gov, search engines, federated search, Google, multilingual

The development of OSTI

Established in 1947, the U.S. Department of Energy (DOE) Office of Scientific and Technical Information (OSTI) (http://www.osti.gov/) fulfills the agency's responsibilities to collect, preserve, and disseminate scientific and technical information (STI) emanating from DOE research and development (R&D) activities.

OSTI was founded on the principle that science progresses only if knowledge is shared. OSTI's mission is to advance science and sustain creativity by making R&D findings available and useful to DOE and other researchers and the public. The OSTI Corollary – accelerating the sharing of knowledge accelerates the advancement of science – takes OSTI's founding principle to the next level.

OSTI's statutory authority is provided in the Atomic Energy Act of 1946 and in several subsequent laws. In the words of Section 982 of the Energy Policy Act of 2005, "The Secretary, through the Office of Scientific and Technical Information, shall maintain within the Department

publicly available collections of scientific and technical information resulting from research, development, demonstration, and commercial applications supported by the Department." [1]

OSTI grew out of the post-World War II initiative to make the scientific research of the Manhattan Project as freely available to the public as possible. On November 17, 1944, President Roosevelt wrote Vannevar Bush, then the Director of the Office of Scientific Research and Development, to request his counsel on how to capitalize on the experience of the United States' R&D war efforts – most of which was done in utter secrecy – in the days of peace to come. Roosevelt asked for guidance on four major points. This was the very first issue he addressed to Bush:

> First: What can be done, consistent with military security, and with the prior approval of the military authorities, to make known to the world as soon as possible the contributions which have been made during our war effort to scientific knowledge?
>
> The diffusion of such knowledge should help us stimulate new enterprises, provide jobs for our returning servicemen and other workers, and make possible great strides for the improvement of the national well-being, (Bush, 1945).

Bush responded to the President's call with the now famous report, *Science: The Endless Frontier*, published in 1945. In it, he articulated the rationale for a robust governmental role in science and presented the blueprint of how that was to be accomplished. In answering the President's first question, Bush advised that the "the lid must be lifted":

> While most of the war research has involved the application of existing scientific knowledge to the problems of war, rather than basic research, there has been accumulated a vast amount of information relating to the application of science to particular problems. Much of this can be used by industry. It is also needed for teaching in the colleges and universities…. Some of this information must remain secret, but most of it should be made public as soon as there is ground for belief that the enemy will not be able to turn it against us in this war....
>
> The Government should accept new responsibilities for promoting the flow of new scientific knowledge…. (Bush *ibid*)

Furthermore, Bush wrote:

> International exchange of scientific information is of growing importance. Increasing specialization of science will make it more important than ever that scientists in this country keep continually ahead of developments abroad….
>
> *The Government should take an active role in promoting the international flow of scientific information.* (Bush, *ibid*)

In 1945, General Leslie Groves, commanding the Manhattan Engineer District in Oak Ridge, Tennessee, mandated that all classified and unclassified information related to development of the atomic bomb be brought together into one central file. Thus, in 1947, OSTI became home to one of the world's most comprehensive collections of energy-related information, with separate operations for classified information, (Vaden, 1992).

Long before the Internet came along, OSTI advanced science by making research information widely available. OSTI annually responded to upwards of 50,000 requests for information and during the 1977 "energy crisis" fielded more than 150,000 requests. OSTI operated one of the few federal printing plants in the United States, and in 1948 began an almost 30-year production of the world-famous *Nuclear Science Abstracts*, which greatly expanded access to nuclear science information. OSTI shouldered a lead role in providing materials to the Atoms for Peace Geneva Conferences, envisioned by President Dwight D. Eisenhower to pool nuclear information for sharing with peaceful nations. OSTI was instrumental in establishing the International Nuclear Information System (INIS), which promotes nuclear information exchange between 110 countries.

In 1994, OSTI created the first DOE home page, and it has made significant strides into the Information Age ever since, defining new electronic exchange formats, creating collections of digitized scientific and technical information, serving researchers directly, and developing an energy science and technology virtual library. OSTI today hosts three major collections of scientific and technical information: Science Accelerator, which features DOE R&D resources; Science.gov, which provides access to STI from federal science agencies through the U.S. government; and WorldWideScience.org, which offers resources from more than 60 nations around the world.

OSTI has championed an aggressive effort on a series of fronts to make authoritative science information more efficiently available to researchers and the public alike. It has played a leading role in developing and adopting cutting-edge web tools such as relevancy ranking, technology that allows search results to be returned in a ranked order relevant to the search query, and federated search, the simultaneous search of multiple web databases in real time via a single search query, to enhance the diffusion of scientific knowledge.

Along with other federal science agencies that are pioneers in this area, OSTI believes, as former Director of the National Institutes of Health Dr. Elias Zerhouni has put it, that "the real value [in the explosive growth of scientific knowledge] is in the full *connectivity* of *all* available electronic sources of scientific information and their efficient exploitation with the powerful emerging software tools of specialized search engines and not in just posting articles for passive display.", (Zerhouni, 2008)

Back in the old days, OSTI's flagship product, *Nuclear Science Abstracts*, provided abstracts of documents and journal articles together with information about where the document or article could be found, (Vaden, 1992). To be of the most value, the customer had to obtain the full-text document or journal article, but the technology of that day meant that customers were left to their own devices to obtain the full text. Typically, only users on the premises of a large university

library or other major library could access full text. Being able to visit and use such a facility was a privilege available to only a small number of people.

The situation today is starkly different. Today, the typical user of an OSTI product has immediate access to full text. No longer need the user be on the premises of a major library. All he or she needs is internet access anywhere in the country, even around the world. This includes the researchers that DOE funds at hundreds of colleges and universities. It includes the tens of thousands of researchers who use DOE facilities, the million working researchers and tens of millions of students in America, and many more millions around the world.

The upshot is that OSTI is providing cutting edge science to millions of people for whom such privileged access would have been impossible just 15 years ago. And prospects for the future are even more mind boggling. Thanks to the pioneering work of OSTI, the stage is set for the world's cutting edge science and technology to reach billions of people, rather than "mere" millions. And the depth of the science and technology they will enjoy will exceed even today's accomplishments.

The OSTI Corollary and The Knowledge Investment Curve

As Isaac Newton said, "If I have seen further, it is by standing on ye shoulders of Giants."(Newton, 1675). Newton was not alone on those shoulders. Everyone in science, from his day to ours, draws on the work of others.

Science is all about the flow of knowledge: new methods, instruments, techniques, concepts, results, questions, data, etc. The flows are endless, complex and in all directions. It is rightly called a diffusion process. This concept is reflected in a host of statutes that form the legislative basis for federal scientific information agencies such as the U.S. Department of Energy Office of Scientific and Technical Information.

Given that the diffusion of knowledge is central to science, it behoves us to see if we can accelerate it. We note that diffusion takes time. Sometimes it takes a long time. Every diffusion process has a speed. The OSTI thesis is that speeding up diffusion will accelerate the advancement of science.

Innovation has often been linked with prosperity and growth and consequently, trying to understand what drives scientific innovation is of extreme interest. Identifying sets of population characteristics, factors and mechanisms that enable scientific communities to remain at the cutting edge, accelerate their growth, or increase their ability to re-organize around new themes or research topics is therefore of special significance. Yet generating a quantitative understanding of the factors that make scientific fields arise and/or become more or less productive is still in its infancy. This is precisely the type of knowledge most needed for promoting and sustaining innovation. Ideally, the efficient and strategic allocation of resources on the part of funding agencies and corporations would be driven primarily by knowledge of this type.

Every scientist knows that science advances only if knowledge is shared. Mathematically, this statement implies that the advance of science is a function of both the sharing of research results, as well as doing the original research. In principle, therefore, decision makers face the problem of deciding how much to spend on original research and how much to spend on sharing the knowledge that comes out of research.

Consider the graph below (Fig 1) with the x-axis being the fraction of research resources expended on spreading knowledge. The scale would range from 0% to 100%. The y-axis is the pace of scientific discovery.

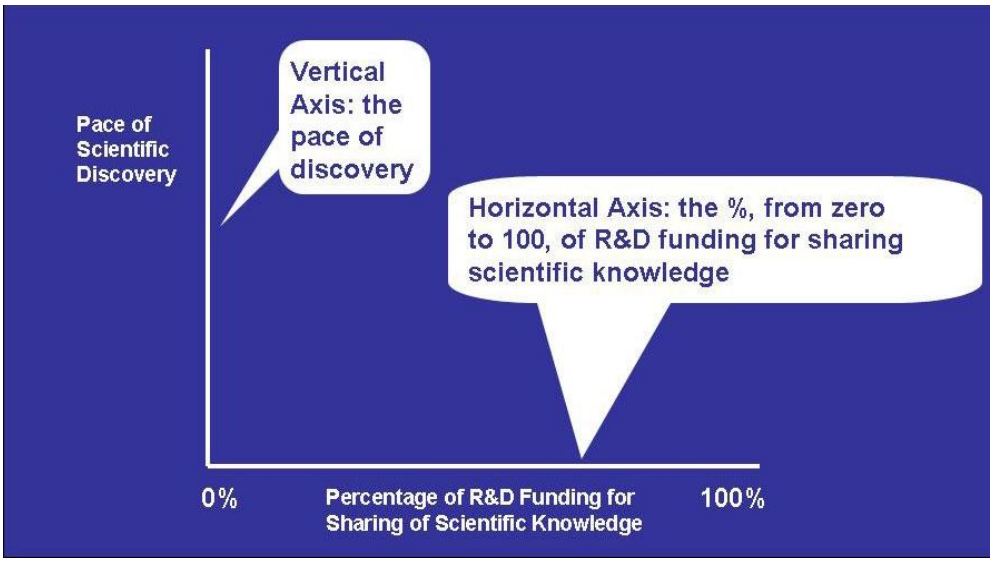

Fig 1 – Relationship between funding for sharing and the pace of scientific discovery

One can imagine a curve plotting the pace of discovery as a function of the fraction of resources expended on sharing knowledge.

When the fraction of resources is 0%, the pace of scientific advancement is zero, as nothing is shared, (Fig 2).

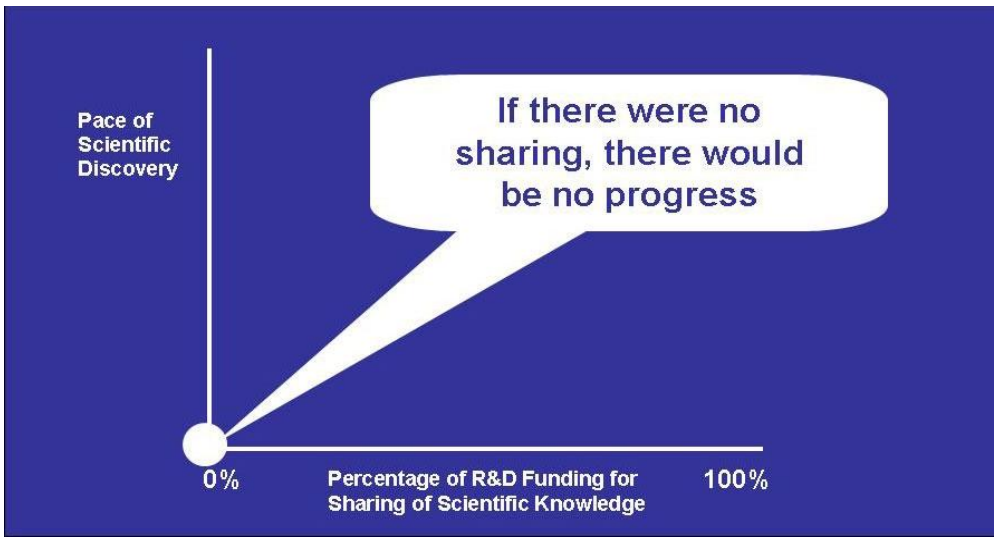

Fig 2 Zero funding for sharing

When the fraction of resources is 100%, the pace of advancement is also zero, as nothing is spent on the research itself, (Fig 3).

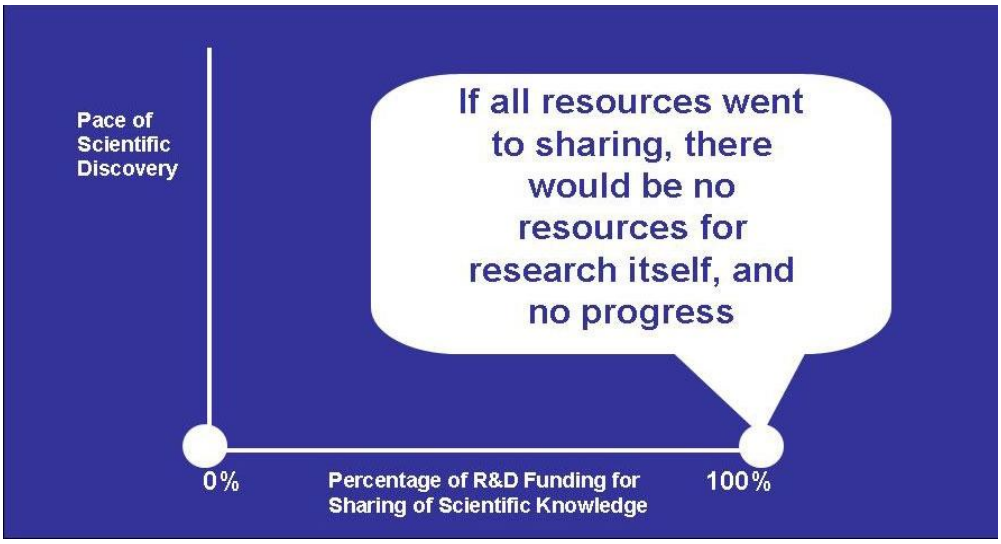

Fig 3 – 100% funding for sharing

In between these endpoints, the plot will have a maximum. The plot is the "Knowledge Investment Curve."

While we show a conceptualization of the Knowledge Investment Curve, we know very little about the actual form of this curve, or even how much is currently invested in sharing.

Most knowledge sharing activities are not funded directly as budget items. These include writing an estimated one million research papers and reports a year worldwide, as well as finding and reading them. It includes preparing for and participating in conferences, as well as writing and reading emails, blogs, and the like. It also includes training postdocs and Ph.D. students, plus an untold number of colleague-to-colleague conversations.

These myriad activities are centuries old, as old as science itself. What each costs in the aggregate we have little idea. We do know that scientific journals cost several billion dollars a year, because they depend on a central infrastructure that has a visible price. We also know the budgets of organizations whose purpose is to share knowledge such as the DOE Office of Scientific and Technical Information and sister organizations across the U.S. government such as the Defense Technical Information Center, the National Library of Medicine, the National Agricultural Library, and others.

We also know that the Internet, especially the World Wide Web, is changing the nature of the equation, because the unit cost of sharing is so much less than the traditional means. The web has made sharing global, or at least potentially so.

We can ask then what the federal investment in web-based science sharing should be. Conceptually, points on the Knowledge Investment Curve to the left of the optimum imply that the pace of science discovery would be accelerated by increasing the percentage of funding for sharing results, Fig 4). One thing we know is that the investment in sharing is highly uneven across the various sciences. The fraction of health science research funding dedicated to sharing knowledge is greater than for physical and energy sciences. The latter is unlikely to be near the optimum.

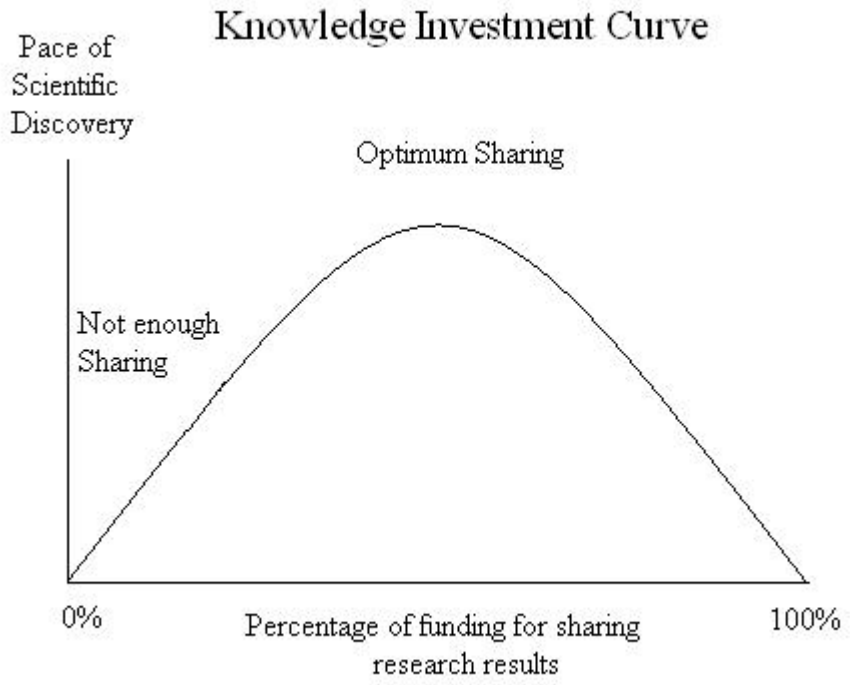

Fig 4 Knowledge investment curve

Information sharing is an integral part of the R&D process. Thus, decision makers affect the pace of scientific progress when they determine the fraction of R&D dollars dedicated to sharing knowledge. Think of it this way: A program for sharing knowledge derived from scientific research has much in common with a scientific research program itself in that they share the common goal of advancing science. When decision makers of R&D programs discuss optimum funding for research, their decisions are driven by affordability. Similarly, there is an optimum investment in sharing research results as conceptually suggested by the Knowledge Investment Curve. And just as for research itself, the optimum investment is not the minimum.

The OSTI Corollary – If the sharing of knowledge is accelerated, discovery is accelerated – explains why we at OSTI are constantly striving to share more science with more people faster and more conveniently than ever before.

Much of Science Is Non-Googleable

Before we can accelerate the sharing of knowledge, however, we must dispel the misperception that traditional search engines are already doing the job.

Basically, there are two ways to get to knowledge on the web. In fact, one can think of them as two kinds of knowledge. The first is the ordinary web page, of which there are several billions.

This sea of web pages is what is searched when you use a search engine like Google. We call it the "surface web."

Google, Yahoo!, Bing, and other conventional search engines do for the surface web what publishers have long done for books – they create an index so that customers can quickly find information. Web users value this service so highly that search companies have become phenomenally successful enterprises.

However, beneath this surface web there are vast document repositories. They often have their own search tool for searching within the repository, but traditional search engines like Google do not reach within these databases, even though they are web accessible. We call the part of the web in which repositories and databases reside the "deep web."

The deep web is huge. By some estimates, it is more than 500 times the size of the surface web. Analysts estimate that perhaps 99 percent of all the web-accessible scientific documents are in deep-web databases. Because these documents are not accessible to search engines and robots, this creates a huge gap in knowledge searchability, (Bergman, 2000). In fact, much of the information on the web is inherently unavailable to Google and Yahoo! This key limitation would come as a surprise to many web users. The concept that if you "Google" long enough you can find it is so firmly entrenched in the web-cognizant public that the word "Google" has been elevated to a verb. In fact, at its annual meeting early in 2010, the American Dialect Society selected "Google" as its "word of the decade," noting it is on the verge of replacing the verb "to search.", (American Dialect Society, 2010). This has led folks at OSTI to do some word-creation of our own. It naturally follows that the adjective derived from that verb is "Googleable," referring, of course, to information that can be found by "Googling." It is just a short jump to arrive at the antonym "non-Googleable," referring to information that cannot be found by Google.

Google founder Larry Page delivered a speech at the 2007 annual meeting of the American Association for the Advancement of Science (AAAS) in which he noted that much of science is not available for Google to retrieve, (Page, 2007). The July 27, 2007, issue of *Science* presented an article by a Google research director who acknowledged the same thing. (Henziger, 2007). I coined the term non-Googleable because the concept is so critically important to science. It turns out that great quantities of science knowledge are non-Googleable. This observation is profoundly important for science in general – and OSTI in particular.

The limitations of traditional web-searching are inherent in the underlying technology used by Google and by each of the other conventional search engine companies to index the web. To get ready for searches by web patrons, a web crawler (or "spider" or "robot") visits many web sites, mostly by following links. An index of each such site is thus created, slowly building a vast composite index of all the sites visited. Later, when web patrons perform a search, they are actually searching the composite index. Difficulty arises because vast numbers of web pages cannot be accessed by following links. In other words, such web pages are not crawl-able. For example, to find an e-print on a database of e-prints, it is typically necessary to enter a search

term on the front page of the database. At this point, a crawler is stumped. As a consequence, the content of the database is not accessed by the crawler, and that content is non-Googleable.

Conventional search engine companies recognize this problem and acknowledge that they cannot solve it alone. Rather, they encourage database owners to take special steps to accommodate their crawlers. Some owners take such steps, others do not. But the root cause of the problem is that crawling technology is inherently limited.

Asking a scientist, engineer, or educator to find information in their field using common web browsers is like asking a doctor to diagnose disease without X-rays, MRI, or any other piece of diagnostic equipment.Information in the deep web can only be mined for data using search engines designed for that particular database. Many of the search engines that are available to mine databases often do not use relevance ranking, making filtering through the information risky and uncertain. Under the current system, finding information in the deep web is a series of practical impossibilities, frustrating internet users, especially scientists and science educators.

Clearly, the web is a transformational technology for sharing knowledge. It's like the Model T Ford – revolutionary but ready for vast improvement. This is especially true when it comes to making the web work for science and technology.

Important lessons for students of web evolution can be learned from examining previous transformational technologies, like the railroads in the 19th century or the automobile in the 20th century. Just as the web first fully captured public imagination in 1994, the automobile first captured public imagination about 1903, when Henry Ford introduced his first mass-produced car. In the years that followed, the Ford made very significant technological progress. Some sixteen years later, Ford offered electric lights, secure doors, a roof, and numerous improvements under the hood, (Ford Motor Company, n.d). Similarly, 16 years after the emergence of the web, considerable progress has been achieved, such as surface web search engines and relevance ranking.

As we all know, Ford's technology was not static in 1919. The automobile continued to evolve, and indeed continues to evolve today. Likewise, the web continues to evolve, even though the specific features of that evolution are currently unknown.

At OSTI we can cite several examples where we have been on the leading edge of technological development. In some cases, we have been on that leading edge in cooperation with some of the largest, most innovative technology companies in the world.

For example, OSTI has been on the forefront of the development of the Site Map Protocol, an open standard for web sites that allows search engines to readily identify the location of all pages on the site, including database records lying behind a search form. In 2003, OSTI began working with Google and Yahoo! to make the research and development findings of the U.S. Department of Energy available through their search engines. As this work progressed, OSTI provided feedback and worked closely with the technical teams from both companies to ensure that our collection was indexed and made retrievable. This work, along with other parallel

efforts by the search engine giants, eventually led to Google unveiling the Site Map Protocol in 2005. OSTI was a first adopter of this technology, which was no great leap as our early work was part of the basis for this standard, which has has been embraced by Google, Microsoft, Yahoo, Ask.com and others. As a result, any government site using this standard can reach Americans through all major search engines, (Google Public Policy Blog, 2007).

We may not be able to predict what the web will look like in a few years, but we can learn lessons from history. Ford made the automobile more user-friendly, faster, easier to operate and most importantly, Ford made the automobile ubiquitous. Just as Ford sensed the burden of transportation was an obstacle to human progress, we know that the burden of searching is an obstacle to science progress. Ford transformed the behavior of the traveling public.

Google is capitalizing on this early era of web technology and is hugely successful, powering more than half the world's searching. We are just in the beginning of this transformation and it will be fascinating to watch and participate in the evolution of the search technology of the future.

In fact, a new, promising technology is now emerging: federated search.

Federated Search

As search capability is key to OSTI's mission, the limitations of crawling have motivated us to find another way to make information in multiple databases searchable. It is called federated search, and it drills down to the deep web where scientific databases reside. OSTI has been a pioneering force in federated search technology since the late 1990s.

Federated search allows users to search multiple data sources simultaneously, in parallel, using a single query from a single user interface. OSTI offers federated search to patrons as a free aggregator of multiple government R&D-related databases.

Here is how federated search works. A web patron seeking science information opens a portal search tool like Science.gov and enters a query, just as he or she would do at Google. But, while the patron's experience looks like Google, the architecture behind federated search is entirely different. The query is transmitted to a central server – in Oak Ridge, Tennessee, in the case of Science.gov and WorldWideScience.org – and then it is fanned out to each of a suite of databases geographically spread out across the U.S. or even the entire world. At each database, the query causes a search to be executed and produces a hit list of search results summaries which might include title, author and snippet. The hit list is then transmitted back to the central server, where the hits are relevancy ranked and sent on to the web patron. In the span of about 20 seconds, the query has been transmitted to numerous databases, searches executed at these databases, and the results brought back and ranked for the patron.

Along the way, OSTI has taken every opportunity to encourage the rapid maturation of federated search technology. Most notable was the development of relevance ranking in a federated environment. Before relevance ranking, federated search results were presented in long lists: a

set of hits from source A would be followed by a set of hits from source B, and then from source C, and so on. Soon the patron was overwhelmed with sets of hits. As with surface web search engines, like Google, relevance ranking was a major advance in meeting the needs of patrons. The challenge was that the technology behind relevance ranking for Google does not work in a federated environment. So new relevance ranking had to be invented.

Federated search is inexpensive to implement and allows for fielded searching, which provides users who know very specifically what they're looking for (e.g. author, title, or publisher) the capability to perform a precision search.

Federated search technology is of particular strategic value to OSTI in that it does not place any requirements or burdens on owners of databases. This means that when an agreement is made with a scientific organization to make its content searchable by one or more OSTI applications, setting up access to the organization's content is a rapid and straightforward process. If the organization's content is already searchable, via the web or some other mechanism, then the organization has no responsibility other than to keep its database accessible, a responsibility it already has.

Another great value of federated search is that databases can be aggregated into federations of federations. This means that a federated search application can act as a single database to another federated search application.

Having layers of federation provides two tremendous benefits to OSTI. First, it greatly extends the reach of a single application from several dozen databases to literally hundreds of databases in real time. This ability to scale is critical to OSTI's drive to accelerating the diffusion of science. Second, multi-layered federation allows for managing collections of content databases in a decentralized way. While it would be too onerous a task for a single organization to manage the availability of hundreds of databases, it is quite manageable for several organizations each to manage access to smaller sets of databases and to provide access to the databases they steward through a federated search application which they then provide as "feeds" to the larger application.

Science.gov

OSTI conceived and hosts Science.gov (http://www.science.gov/ver5.html), a gateway to U.S. government scientific and technical information. Science.gov was launched in December 2002 and pioneered the use of federated search within the federal government. It is an interagency initiative of 18 U.S. government science organizations with 14 federal agencies that contribute content to serve the information needs of the science-attentive citizen, including science professionals, students and teachers, and the business community. It is in its fifth generation and offers access to more than 40 databases and 200 million pages of science information via a single query.

Science.gov is among 10 government websites already “meeting and exceeding” the Obama Administration’s transparency goals, according to a special report by *Government Computer*

News, released July 27, 2009. "Great .Gov Web Sites" described "10 sites that take government to the next level," noting that Science.gov "breaks down stovepipes of research.", (Jackson, 2009) Science.gov is the U.S. contribution to WorldWideScience.org which is described below.

In April 2007, OSTI introduced Science Accelerator, utilizing the proven Science.gov federated search architecture, (http://www.scienceaccelerator.gov/). It searches 10 key DOE resources, including the results of DOE's R&D projects and programs, descriptions of R&D projects under way or recently completed, major R&D accomplishments, DOE patents, and recent research of interest to DOE. Several of these DOE databases maintained by OSTI are now featured as "National Assets" on Data.gov (http://www.data.gov/list/nationalassets). In addition, Science Accelerator is the DOE contribution to Science.gov.

What is particularly interesting about Science Accelerator is that a number of its resources are themselves federated search applications. Thus, Science Accelerator demonstrates the feasibility of building federated search applications hierarchically, where one searchable database is aggregated from multiple searchable databases, each of which can be decomposed further into searchable databases, and so on. This hierarchical construction will allow Science Accelerator to scale to search at least 1,000 databases in parallel in the foreseeable future. This will have the remarkable effect of enabling users to search all web-accessible collections of scientific knowledge related to the DOE mission from a single search form.

WorldWideScience.org

WorldWideScience.org (http://worldwidescience.org/) is a brand new global science gateway that relies on federated search.

In June 2006, at the annual conference of the International Council for Scientific and Technical Information (ICSTI), I proposed a vision for a "Science.world," a global extension of the national model of Science.gov. This was an idea I had conceived and discussed with OSTI and DOE staff in 2005.

Later in 2006, the British Library expressed a desire to partner with the U.S. Department of Energy to develop this international federated search application, and on January 21, 2007, then Under Secretary of Energy Dr. Raymond L. Orbach signed a bilateral statement of intent with the Chief Executive of the British Library, Dame Lynne Brindley. The two officials invited other nations to join in this partnership.

Five months later, at the 2007 ICSTI conference in Nancy, France, OSTI debuted the global science gateway we had named WorldWideScience.org. At the time, WorldWideScience.org performed federated searching of 12 databases and portals across 10 countries. Science.gov was the U.S. resource searched by WorldWideScience.org.

With the successful launch of WorldWideScience.org and the resulting publicity, several other countries approached OSTI seeking to have their science and technology databases added to the global portal. In addition, discussions ensued among participating countries and ICSTI to

transition the bilateral (U.S. /U.K.) governance of WorldWideScience.org to a multilateral structure, and a Terms of Reference governance document was drafted.

In February 2008 at the ICSTI winter meeting in Paris, the Terms of Reference were ratified, defining the purpose, objectives, terms, conditions, and structure for a WorldWideScience Alliance. The Terms of Reference provided that OSTI would serve as Operating Agent for WorldWideScience.org and secretariat to the Alliance. In addition to its member organizations representing various countries, the Alliance would be closely affiliated with ICSTI.

We officially launched the WorldWideScience Alliance on June 12, 2008 at the annual ICSTI conference in Seoul. The launch marked a key milestone, where organizations representing 38 of the 44 countries agreed to take part in the governance and funding of WorldWideScience.org. Our Korean counterpart organization, the Korea Institute of Science and Technology Information (KISTI), hosted a WorldWideScience Alliance signing ceremony, where then DOE Associate Under Secretary for Science Jeffrey T. Salmon congratulated Alliance members and predicted that WorldWideScience.org "will become . . . the Alexandria Library of the 21st century." He emphasized the underlying goal of WorldWideScience.org in stating that ". . . spreading scientific facts and ideas will speed up the pace of discovery."

On October 14, 2008, OSTI announced that the People's Republic of China had joined the WorldWideScience Alliance.

By any metric, WorldWideScience.org is growing at a powerful rate.

- When first introduced, WorldWideScience.org included only 10 countries and 12 databases and portals, and it represented roughly 12% of the world's population.
- A year later, when the Alliance was formally established, on June 12, 2008, the number of member countries had increased more than four-fold, to 44.
- The 38 nations that were represented in the Alliance's founding document, plus 6 others, contributed 32 databases and portals, and represented roughly 53% of the world's population.
- Today, 61 countries contribute 61databases and portals and represent more than 75% of the world's population, enabling DOE and other U.S. scientists to search these sources with a single query. Results are then collectively ranked in relevance order. Features such as alert services enable scientists to stay abreast of ongoing research in their fields, regardless of international boundaries.
- When WorldWideScience.org launched, it provided searchable access to roughly 200 million pages of science content; today it searches across about 400 million pages of important scientific portals worldwide. That's a lot of science information accessible from one search box – equivalent to a shelf of documents 20 miles long.

This is the first time of which we are aware that federated searching has been accomplished on a global scale.

Without WorldWideScience.org to search the national portals, information customers faced a task so forbidding that it was a practical impossibility. Without WorldWideScience.org, customers would have to overcome three formidable roadblocks. First, to search individual national portals, they had to know that those portals exist. We have yet to encounter anyone who knew more than a few such portals. This point may resonate with those in the interlending and document supply community, as their customers must first know what is available to be loaned and supplied.

But let's magically assume that roadblock away. Let's assume for a moment that the information customer somehow knows about all 61 national portals. Then the customer would face the second formidable task of visiting each portal and searching it one by one. The customer would face a daunting task.

But let's magically assume this roadblock away, too. Let's assume for a moment that the information customer did visit and search each portal. Then, the customer would be faced the third roadblock of sorting through 61 long hit lists – yet another imposing task.

Thus, WorldWideScience.org changes a practical impossibility to an easy and rewarding function by searching portal upon portal of science information typically not searched by conventional search engines, in parallel, with only one query, ranking the results, and thus saving tremendous time and effort. So where once we had isolated portals of information, we now have portals working as a unit, an integrated whole. Federated search, through a gateway such as WorldWideScience.org, speeds communication, accelerates discovery, and expedites scientific and economic progress. And to use WorldWideScience.org, all that is needed is internet access from anywhere in the world.

Of course, most researchers have at least one, if not more, commercial databases they regularly search for their information needs. Researchers often have a set amount of time they are willing to devote to literature searching and, thus, want to spend their time on the most productive databases. A federated search tool that returns productive results from 61 portals that normally would not be searched is a tool that is worthy of some of that valuable search time.

Many of the databases searched through WorldWideScience.org are not well known outside their originating countries and are not easily accessible through typical commercial search engines. In fact, a recent analysis indicated that WorldWideScience.org results, when compared to Google and Google Scholar results, were unique approximately 96.5 % of the time.[2] WorldWideScience.org makes it easy for DOE and other U.S. researchers to find and search these sources.

Despite these remarkable advances, there is more work to do. Currently, WorldWideScience.org is limited to searching databases with English titles and abstracts. This constraint confines the number of databases accessible by the search engine.

The Alliance is now exploring translation technologies to expand the network of databases accessible to the worldwide community and is making progress toward deploying this capability.

A prototype allows users to select their preferred language. Queries are translated into the languages of the databases being searched and results are then returned in the user's language.

We are committed to launching Multilingual WorldWideScience.org at the ICSTI meeting in Helsinki in June 2010. With the advent of multilingual translation capabilities, databases in other languages, such as Russian and Chinese, will be available to U.S. scientists. Multilingual translations will provide DOE and other U.S. researchers with access to vast amounts of non-English scientific information. Conversely, Russian and Chinese scientists will also be able to search English databases in their native languages and receive search results translated into Russian and Chinese. The same functionality will be available to speakers of all the world's major languages.

The importance of sharing science knowledge is not new, but its realization, even in the Information Age, had not been possible on such a large scale until the development of WorldWideScience.org. And, while WorldWideScience.org will constantly seek to improve with new features, sources, and functionality, today it represents a groundbreaking development in access to global science resources. The broad participation in the Alliance indicates that a similar view is shared by many countries.

A Billion Pages of Authoritative Science?

For more than a decade, OSTI has been leading the charge related to federated search technology, specifically applying this technology to enhance access to scientific and technical information from government science research agencies at home and abroad.

OSTI, through federated search, ensures access to non-Googleable science. In fact, through OSTI products, librarians, researchers and the public can access a science page count comparable to, but not duplicative of, Google's entire science content, (Fig 5).

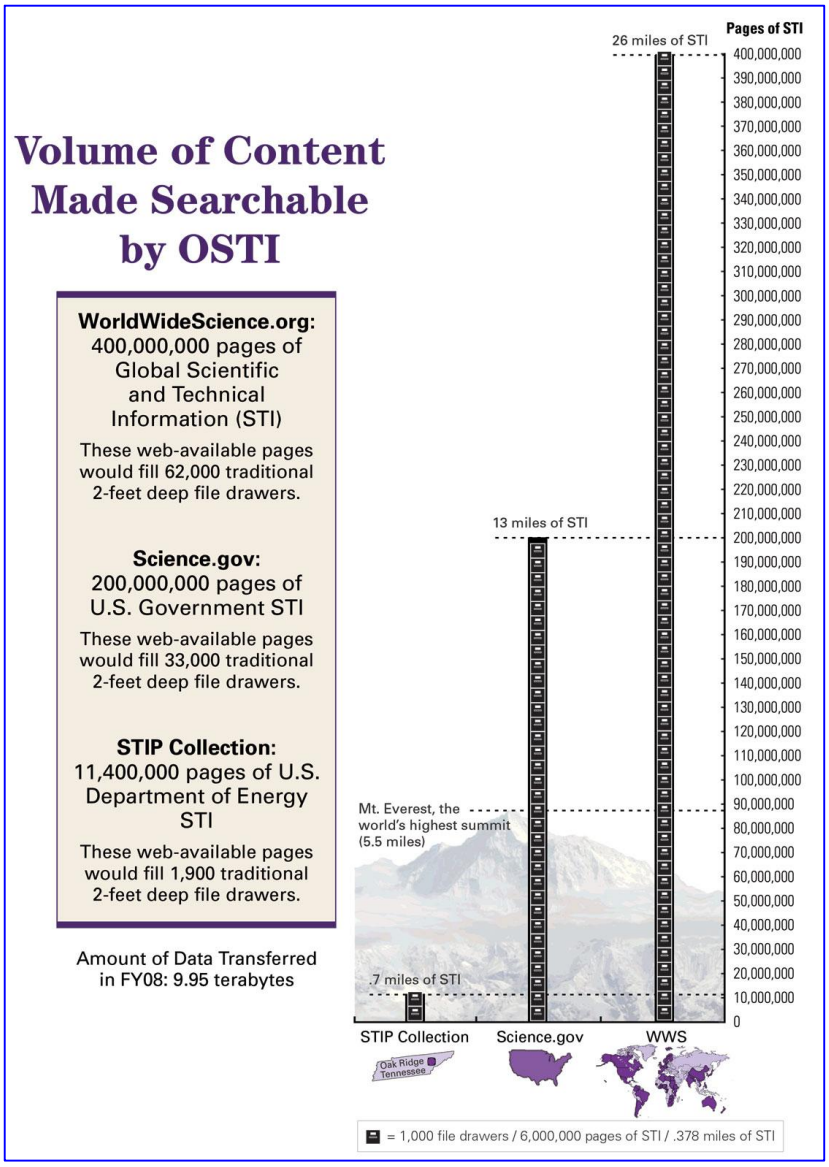

Fig 5. Content made searchable by OSTI

The model that Science.gov introduced and that Science Accelerator and WorldWideScience.org have propagated has proven itself. While as far as the eye can see, there will be a place for crawling and indexing, federated search may one day grow to the point that it becomes the dominant search architecture in the deep web.

For the serious researcher and the science-attentive public, the future of federated search is bright. OSTI was a pioneer in introducing federated search into the federal government, and OSTI continues to pioneer innovative approaches to managing and overcoming the challenges to quickly delivering the most relevant high quality content. We at OSTI are ready to scale up our efforts in federated search which has advanced very rapidly over the last few years and should continue to do so. However, neither crawling nor federated searching is a panacea. Federated

searching does things that crawling can't do, and crawling does things that federated searching can't do – they are complementary technologies. Portals like WorldWideScience.org, which make non-Googleable information searchable via federated search, are complementary to conventional search tools like Google and Yahoo! that rely upon crawling.

Federated search technology is not without its challenges, the greatest of which is the cost of building and maintaining the software "connectors" that search and retrieve documents from deep web databases. The increasing adoption of standards among content publishers' search interfaces is driving down the cost and effort to build and maintain connectors.

The other great challenge of federated search, speed of obtaining search results due to performance limitations at the content providers' sites, is being addressed through OSTI-funded research including caching of common search results, automatic selection of the right sources to search to minimize unnecessary load placed on sources not relevant to a query, strategic mirroring of content, and other approaches.

The future

What is next? There is no inherent reason that a single tool cannot rely upon both a crawled index and a live federated search in parallel. Indeed, OSTI's largest product does just that. It is the E-print Network (http://www.osti.gov/eprints/). All in parallel, it searches 1.5 million e-prints that have been crawled, plus an addition 5 million e-prints hosted in 50 e-print databases, comprising in all about 100 million pages. As far as we know, there is no other tool in the world that virtually integrates such a quantity of e-prints. Further, we are not aware of another publicly available search tool that searches federated databases and crawled indexes in parallel.

The crawling done by the E-print Network is different from that done by conventional search engines. The E-print Network crawls only those sites of known quality. Such filtering produces a high quality search tool. There would seem to be great potential to build on this theme of combining into a single information product searches of crawled indexes and federated search of databases.

Here is one potential application. It would be technologically possible to combine WorldWideScience.org and crawled indexes. In addition, it would be technologically straightforward to add in more federated search tools to make an enormous search tool. The builders of this "uber" tool of the future would need to be careful about relevance ranking, but that challenge is manageable.

Another opportunity in the near term is for private-sector organizations to take advantage of government science federations and integrate them with proprietary content. Thus, the combination of crawled indexes and federated searches is an extremely promising path to the future. A billion-page, high quality science search tool may be available soon to spread ideas, increase learning, and further accelerate the progress of science.

Yet making more information available is not enough. It must be presented more conveniently, so that it is easier and faster to find. To this end, another key to success is precision searching.

The problem that is inherent with federated search techniques is information glut. Customers can get so many hits that it is beyond human capacity to sort through them. The way around this problem is with precision searching, one version of which is relevance ranking. We are all familiar with relevance ranking. It is what Google does. Google itself credits its relevance ranking for its success.

But the methodology for relevance ranking on the deep web is far different than on the surface web. To this end, relevancy ranking is being reinvented for federated searching.

Document supply

One of the major challenges to users of Science.gov and WorldWideScience.org is that many of the citations do not include links to the full-text. Indeed, a great deal of the literature is not available in electronic format. Thus, the problem is how to get copies of documents listed in search results that are not online. Some of the individual databases or portals attempt to help with this by offering options to purchase copies, or providing links to publishers' websites, Google Books, or other information providers. However, there are still many citations that offer no clue how to obtain the cited documents

OSTI has recently been working on adding links in the Energy Citations Database (http://www.osti.gov/energycitations/) to www.worldcat.org. WorldCat is a database containing records for over 10,000 libraries worldwide and over 1.4 billion items available in those libraries. Searchers will be able to click on the WorldCat link and get a list of nearby libraries that hold the desired document. Even if no local library holds the document, WorldCat will show where document can be obtained. In the future we will be looking at providing similar links on other OSTI databases and federated search products. Providing such links in a federated search product will be a tremendous challenge, but should be extremely helpful to searchers. Whether by print or by pixel, OSTI long has been committed to ensuring appropriate and ready access to government research. As a leader in making the web work for DOE science, OSTI is embedded in the Internet transformation, and OSTI itself is being transformed. Our dual core mission – getting DOE results out to the scientific community and beyond, and getting the community's results into DOE – has not changed. But the technology we apply to that mission has changed a lot. By carefully adopting Internet technology, and even pioneering new advances in that technology to meet our needs, OSTI achieves its mission better than ever before.

Conclusion

OSTI is dedicated to the principle that, to advance science, research must be shared. OSTI works to accelerate discovery by speeding access to knowledge and it is transforming the behavior of the research scientist. OSTI is doing everything possible to make use of the evolving Internet to diffuse knowledge related to our agency mission.

Simply put, we at OSTI intend to make more science accessible to more people more conveniently than has ever been done before.

Footnote

[1] Statutory authority for the responsibilities of the Office of Scientific and Technical Information has been codified in the enabling legislation of the U.S. Department of Energy and its predecessor agencies:

- Atomic Energy Acts of 1946 (P. L.79-585) and 1954, as amended (P.L. 83-703) established a program for the dissemination of unclassified scientific and technical information and for the control of classified information (42 U.S.C. Sec. 2013, 2051, and 2161).
- Energy Reorganization Act of 1974 (P.L. 93-438) defined responsibilities for developing, collecting, distributing, and making scientific and technical information available for distribution (42 U.S.C. Sec. 5813, 5817).
- Department of Energy Organization Act of 1977 (P.L. 95-91) provided for maintaining a central source of information and disseminating information (42 U.S.C. Sec. 5916, 7112).
- America COMPETES Act of 2007 (P.L. 110-69), Section 1009, required that Federal agencies that conduct scientific research develop agency-specific policies and procedures regarding the public release of data and results of research.
- Energy Policy Act of 2005 (P.L. 109-58), Section 982, specifically called out OSTI's responsibilities.

[2] This finding was produced by OSTI in a comparison of Google, Google Scholar and WorldWideScience.org search results conducted in August 2009. OSTI scientists constructed 33 queries across a wide range of disciplines (chemistry, physics, medicine, etc.). In an effort to mimic typical scientist search behavior, queries had a fairly highly degree of specificity. For some search phrases, OSTI used double quotes to emphasize precision; on others, quotes were not used. Search results from the 33 queries were captured from Google, Google Scholar and WorldWideScience.org. Overlap in the search results sets was identified by searching for exact title matches. Across the 33 queries, WorldWideScience.org results were uniquely different from Google and Google Scholar results 96.5% of the time.

OSTI is preparing a technical paper about these findings and the methodology employed. OSTI previously has reported these findings in the following presentations and article:

- Warnick, W. (2010), "Open Innovation Enabled by Global Networking of Science and Technical Knowledge," Presentation, January 27, Collaborative Expedition Workshop, National Science Foundation, Arlington, VA <http://www.osti.gov/speeches/fy2010/NSF/index.shtml> (Slide #13)

- Warnick, W. (2009), "Federated Search (Emphasizing WorldWideScience.org) as a Transformational Technology Enabling Knowledge Discovery," Presentation, October 20, Interlending and Document Supply Conference, Hannover Germany <http://www.ilds2009.de/> <http://www.osti.gov/speeches/fy2010/IDSC/index.shtml> (Slide # 15)

- Johnson, L. (2009), "WorldWideScience.org: The Importance of Being Unique," OSTI Blog, October 13 http://www.osti.gov/ostiblog/home/entry/worldwidescience_org_the_importance_of

References

American Dialect Society. (2010), "2009 Word of the Year is 'tweet'; Word of the Decade is 'google,'" News release, January 8: (http://www.americandialect.org/index.php/amerdial/2009_word_of_the_year_is_tweet_word_of_the_decade_is_google/).

Bergman, M.K. (2000), "The Deep Web: Surfacing Hidden Value," white paper, BrightPlanet Corporation, Sioux Falls, SD (http://www.brightplanet.com/industry-insight/white-papers/).

Bush, V. (1945); (1960 reprint), *Science: The Endless Frontier*, National Science Foundation, Washington, D.C., p. 3.

Bush. *Ibid.*, p. 8.

Bush. Ibid., p. 22. Emphasis in the original.

Ford Motor Company. (n.d), See "The Model T Put the World on Wheels," Ford Motor Company web site: (http://www.ford.com/about-ford/heritage/vehicles/modelt/672-model-t).

Google Public Policy Blog, (2007), "Senate testimony: Our efforts to better connect citizens to government," December 11, See: (http://googlepublicpolicy.blogspot.com/2007/12/senate-testimony-our-efforts-to-better.html).

Henziger, M. (2007), "Search Technologies for the Internet," *Science,* July 27, vol 317 no 5837, pp. 468-471.

Jackson, J. (2009), "Great .Gov Web Sites 2009," *Government Computer News*, July 27, See: (http://www.gcn.com/Articles/2009/07/27/GCN-Great-Gov-Web-Sites-2009.aspx).

Newton, I. (1675), Letter to Hooke, 5 February 1675/6. In Turnbull, H.W. (ed.) (1959), *The Correspondence of Isaac Newton*, I, 1661-1675, vol. I, 416, as cited in Bynum, W.F. and Porter, R. (eds.) (2005), *Oxford Dictionary of Scientific Quotations*, Oxford University Press, New York, NY, p. 460.

Page, L.(2007), Plenary Lecture, American Association for the Advancement of Science Annual Meeting, February 16, San Francisco, CA (video: http://www.youtube.com/watch?v=8_3OCq_vTWM).

Vaden, W. M. (1992), *The Oak Ridge Technical Information Center: A Trailblazer in Federal Documentation*, U.S. Department of Energy, Oak Ridge, Tennessee, pp. 1-17 (http://www.osti.gov/history).

Vaden, *op. cit.*, pp. ix-x, 46, 77 (http://www.osti.gov/history)

Zerhouni, E. A.(2008), , Presentation to the Subcommittee on Courts, the Internet and Intellectual Property, Committee on the Judiciary, U.S. House of Representatives, September 11, 2008, slide 4. Emphasis in original.

.

Author biography
WALTER WARNICK, Ph.D.

As Director of the Office of Scientific and Technical Information (OSTI) in the U.S. Department of Energy Office of Science, Dr. Warnick directs the agency's scientific and technical information operations. He embraces the opportunities offered by the web to accelerate the spread of knowledge about science and technology. To this end, he has championed aggressive efforts to capitalize on technological advances to develop and provide state-of-the art products and services for sharing knowledge. Dr. Warnick and his colleagues continuously work to further advance web search technology.

Dr. Warnick was elected Fellow of the American Association for the Advancement of Science (AAAS) in 2005 "for leadership in the federal scientific information community and for contributions to the conceptualization, development, and implementation of innovative programs that significantly advance access to government information." He earned his Ph.D. and M.S. in Mechanical Engineering from the University of Maryland and his Bachelor of Engineering Science from The Johns Hopkins University.