

DOE Science Accelerator

Advancing Science by Accelerating Science Access

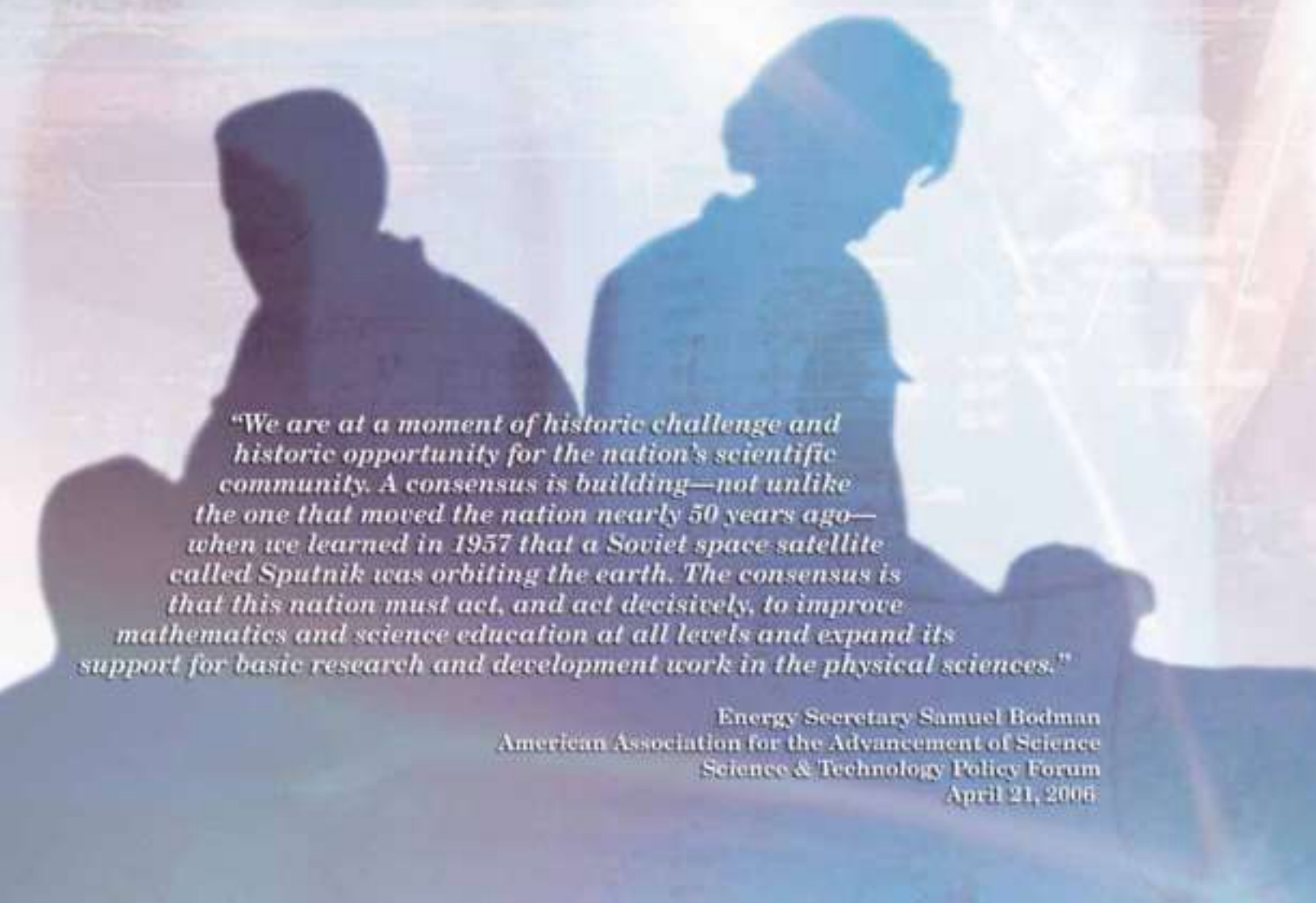


Delivering smart, global search technologies to speed discovery

U.S. Department of Energy
Office of Science
Office of Scientific and Technical Information (OSTI)
June 2006

Table of Contents

The Vision: Speeding Discovery	1
The DOE Science Accelerator	4
Initial Thrust—Education	5
The Broader Agenda: Global Discovery	7
OSTI's Unique Position	9
The Path Forward	11



“We are at a moment of historic challenge and historic opportunity for the nation’s scientific community. A consensus is building—not unlike the one that moved the nation nearly 50 years ago—when we learned in 1957 that a Soviet space satellite called Sputnik was orbiting the earth. The consensus is that this nation must act, and act decisively, to improve mathematics and science education at all levels and expand its support for basic research and development work in the physical sciences.”

Energy Secretary Samuel Bodman
American Association for the Advancement of Science
Science & Technology Policy Forum
April 21, 2006

Foreword: Why We Need a DOE Science Accelerator

To accelerate discovery, it is essential to accelerate the diffusion of science knowledge. This calls for a new era in the sophistication and breadth of the tools to access and use scientific knowledge. Herein, the Office of Scientific and Technical Information (OSTI), an organization of the U.S. Department of Energy (DOE) Office of Science, proposes the “DOE Science Accelerator.”

Why build the DOE Science Accelerator? Because it is impractical for researchers to spend time finding and sifting through hundreds, if not thousands, of information sources in various disciplines and still have time for life-altering discoveries of their own. Scientists and science-attentive citizens need a time-saving single-search interface for the whole of science. They need to explore the deep Web, where specialized databases are beyond the reach of surface Web crawlers such as Yahoo! and Google. They need transformational knowledge-diffusion technologies that enable robust and rapid scientific discovery. The DOE Science Accelerator will meet those needs.

Why now? Because it is now possible to develop the technology, and the foundation has been laid. A significant milestone was achieved in 2002 when Science.gov introduced the capability to search 30 major databases of federal science agencies. OSTI pioneered this effort, but it has taken the cooperative effort of 16 information organizations from 12 executive branch agencies to successfully launch and sustain this authoritative gateway to scientific knowledge. It is estimated that there are as many as 1,000 additional sources of scientific merit throughout the world of university, non-federal and foreign research entities. Information customers will only be able to reap the full benefit of these resources with the help of global search technology. Specifically, to accelerate advances in science and maximize the return on research investment, it is essential to create a global search capability to make these resources searchable and accessible.

But first we must dispense with the popular misconceptions that Web technology is mature and that the Web provides access to all meaningful information. We only have to look at the evolution of other transformational technologies to recognize that we are just beginning to exploit the Web. Alexander Graham Bell could not have envisioned the cell phone; Henry Ford could not have envisioned the hybrid vehicle. Invented by physicists for communicating about physics, the Web surfaced as a tool for posting and viewing static Web pages. While this remains an important application today, a next-generation model, Web 2.0, is emerging.

Web 2.0 envisions that information customers have continuously evolving computer-based services such as relevance ranking and searching support. OSTI pioneered federal government Web 2.0 applications for the public several years before the term was coined—and we have only just begun. Making science resources in the deep Web globally searchable cries out for a Web 2.0 solution.

We are answering that call with the DOE Science Accelerator.

Walter L. Warnick, Ph.D.



Director, OSTI

The inability to globally search the Web is an enormous gap that frustrates the diffusion of science knowledge.

The Vision: Speeding Discovery

The need for discovery calls for a new era in access to science. The DOE Science Accelerator will be the flagship of this new era.

New discovery is required to meet national and worldwide needs for major advances to power our economy, develop energy independence, and protect our environment. But advances in science are only possible if knowledge is shared. Further, accelerating discovery is enabled by speeding the diffusion of knowledge.

Hence, scientists need the technology to access and search—en masse and with precision—all of the important science document databases worldwide.

The DOE Science Accelerator will yield that technology.

Building upon the DOE Office of Scientific and Technical Information (*OSTI*) success in deploying technology that enables search across distributed document databases, the DOE Science Accelerator will develop the capability to search 1,000 distributed databases in parallel. To do this, resources must be marshaled to overcome the technological barrier of applying this capability to large numbers of distributed databases.

While technology has greatly accelerated the availability of scientific information on the Web, the tools and capabilities to search that information have not kept pace. This lag in search technology has created a giant chasm in the Internet where scientific databases reside but cannot be globally searched.

Because a way does not currently exist to search across large numbers of scientific databases with one query, scientists and science educators are blind to an untold quantity of untapped information. Much like the 19th century physician without x-rays or the 20th century Web surfer without Google, today's scientists and science educators cannot fathom the quantity and quality of information they are missing without the DOE Science Accelerator. *(continued on page 6)*

The good news is that, theoretically, today's scientists and science-attentive citizens can access about 1,000 scientific databases. The bad news is that, as a practical matter, no individual scientist or citizen has the capacity to deal with them. This situation cries out for a computer-based solution—the DOE Science Accelerator.

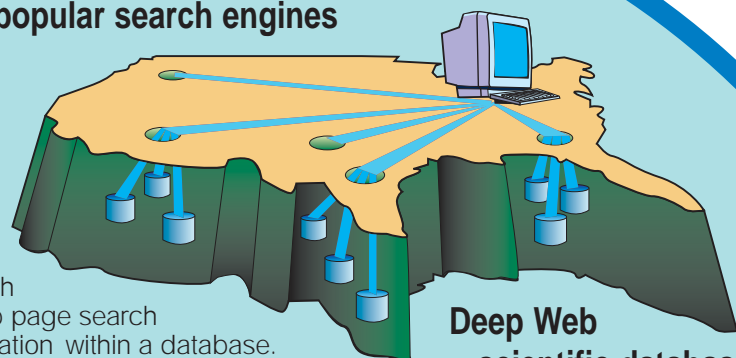
Surface
Web

**Deep
Web**

99% of scientific
documents in databases
are in the deep Web

What is the deep Web and why can't Google reach it?

Surface Web crawled by popular search engines



**Deep Web
scientific databases**

Search engines, such as Google, rely upon automated crawlers and are great for finding Web pages, such as www.osti.gov. However, these Web page search engines typically cannot reach information within a database. Rather, database content is retrieved through the database's own search engine.

For example: On the front page of the database DOE R&D Project Summaries (www.osti.gov/rdprojects/), the database's search box is a doorway to its deep Web content. Here, you can type in search terms and retrieve relevant information from the database. Try this search to prove it: Type "Blue Mountain" into the database search box. The first project returned is entitled "BLUE MOUNTAIN GEOTHERMAL DRILLING - PHASE II DRILLING." Now let's see if traditional Web page search engines can negotiate this pathway. Type "Blue Mountain" into the search box of a traditional Web page search engine. Your results might include greeting cards, ski resorts, and other non-scientific home pages—but no results will describe the R&D project. So we know the project summary is available on the Web, but it cannot be found by traditional Web page search engines.

Recognizing the distinction between searching Web page content and database content is important for science. This is because the bulk of authoritative science information resides in databases within the deep Web, which, as we've just seen, is off limits to Web page search engines. This is why OSTI is proposing to create the capability to search large numbers of databases in the deep Web, making it easier for students, teachers, researchers, corporate R&D labs, and government scientists to find the information they need.

Already, OSTI has pioneered the use of a new class of search engine specifically designed to access distributed resources in the deep Web, enabling a single query to launch searches across a limited number of databases. By using this innovative technology, it no longer matters where the information resides nor what format it is in, and the patron need not know the specific location of the information. While these factors no longer pose barriers to the process of information discovery, a new limitation has emerged—ramping up to larger numbers of databases. The associated technological barriers need to be overcome.

“In science, there is a natural duty to make what is known searchable.”

Kevin Kelly
New York Times Magazine
May 14, 2006

A sampling of deep Web databases

NLM's PubMed: Citations with links to full text
www.pubmed.gov

DOE's Information Bridge: Full-text research documents
www.osti.gov/bridge/

APS Journals
publish.aps.org/

Cornell's e-print archive
<http://arxiv.org/>

USDA's Current Research Information System
<http://cris.csrees.usda.gov/>

NSF's Project Awards database
www.nsf.gov/awardsearch/

EPA Science Inventory
<http://cfpub.epa.gov/si/>

NASA Technical Reports
<http://ntrs.nasa.gov>

AIP Conference Proceedings
<http://proceedings.aip.org/>

DOE R&D Project Summaries
www.osti.gov/rdprojects

Instead, they are left the tedious and time-consuming task of searching “door-to-door” in only the scientific communities and databases with which they are already familiar.

Commercial search engines, such as Google and Yahoo!, crawl across billions of pages of information on the surface of the Web, but they cannot reach into scientific databases (*the deep Web*). Government efforts, such as OSTI's pioneering deployment of cross-database search in Science.gov, have made significant inroads, but there are likely thousands of additional science resources still under-utilized.

While the barrier is large, so is the payoff. The vision of advancing science will, as Secretary Bodman has stated, require vast improvement in math and science education. The DOE Science Accelerator will be an important tool to ensure the Office of Science makes its R&D results readily accessible to speed discovery and raise scientific literacy.

When the DOE Science Accelerator becomes available, scientists and science educators will no longer need to identify and search one-by-one every database useful to their particular knowledge-discovery pathway. Instead, our nation's innovators will have the capability to search all the important science databases useful to the physical sciences via one query. This will illuminate often obscure databases and speed access to scientific information, which will in turn increase the probability of further and more rapid innovation and discovery. We call this “global discovery.” The end result is the acceleration of scientific advancement.

The DOE Science Accelerator

The DOE Science Accelerator is an initiative to accelerate the diffusion of research results, thereby accelerating innovation and the advancement of science.

The DOE Science Accelerator is a search capability that will, for the first time, consolidate and expose to distributed search all of the important Web-accessible collections of scientific knowledge in every scientific community.

The DOE Science Accelerator is an innovation engine that will drive state-of-the-art knowledge tools and technologies, including implementation of grid computing and data discovery to make global discovery a reality.

The Mission

The mission of the DOE Science Accelerator is to advance science by accelerating science knowledge diffusion, using innovative tools and resources to speed access to R&D results and educational resources.

The Key Objectives

- **Global Discovery:** Develop integrated, full-text search of R&D results from disparate scientific communities to better equip scientists to rapidly travel pathways to innovation.
- **Education:** Build and deliver a searchable “gateway” to the nation’s vast but under-used federal government education resources.
- **Collaboration:** Capitalize on multi-agency and institutional open-access initiatives to speed and ease search of and access to text and data and to promote use and re-use of R&D results.
- **Policy:** Promote development of policy and infrastructure for consistent and effective integration and management of textual information and underlying data.

The DOE Science Accelerator will address the urgent need for (1) a capability for scientists and science educators to search the whole of science with one entry point, and (2) a world-class innovation engine for the testing and application of federated search and other emerging Web 2.0 information science technologies.

Who will benefit? Examples include:

The student who can complete assignments more thoroughly and follow his/her natural curiosity

The teacher who is creating lesson plans in areas where he/she is not expert

The researcher entering a new field

The industrial entrepreneur who needs a solution fast to get the product to market

Initial Thrust—Education

In FY 2008, OSTI proposes to build and deliver a searchable, organized “gateway” to DOE’s educational resources, stratified along logical educational levels with particular emphasis on middle school and high school. OSTI would apply its cutting-edge technologies, demonstrated in products such as Science.gov, to all the specialized education resources residing at the Department’s national laboratories.

Development to be carried out by OSTI would include (a) identifying and building distributed access to education content that exists across the DOE complex and (b) tailoring precision searching to the unique needs of the education community. Science documents, teaching tools, study guides, and other educational resources would be made fully searchable and available to teachers and students in urban and rural areas—wherever the Internet reaches.

The DOE Office of Science would capitalize on its support for hundreds of teachers and students by extending access to hundreds of thousands.

As part of a phased effort, OSTI will first focus on federating searching across existing DOE-produced educational resources. OSTI has proven the value of federated searches of science-related information with Web products such as Science.gov, E-print Network, and Science Conferences. The education pathway, tentatively called the Science Education Network, will facilitate the user’s search by categorizing and segmenting resources along disciplinary lines and by education level (*e.g., grade-level appropriate within elementary, middle, and high schools; undergraduate school; graduate school, etc.*). For example, a 6th grader would not get college-level results.

The Office of Science and DOE produce high-quality science education resources, primarily at our national laboratories. Other science agencies also produce excellent educational materials. However, there is a significant void in making these resources searchable and uniformly accessible. This void undoubtedly causes under-utilization of these resources and, consequently, a diminution of DOE’s and the federal government’s impact on science, technology, engineering, and mathematics (STEM) education.

This initiative will fill the existing void and significantly contribute to U.S. strategic efforts and education goals. The effort would not place any unfunded burden on DOE labs to modify or manipulate their education resources.

The mid-term focus will expand federated search across all federal science agency education resources.



“ . . . just as NASA inspires school children with the excitement and beauty of space sciences, just as NIH similarly reaches out to schools to explain the frontiers and the benefits of the life sciences, so should DOE use its vast frontier technological facilities and the collaboration of scientists from all over the world to inspire students and teachers with the rich frontiers of the molecular, atomic, nuclear and sub-nuclear worlds. The Department’s Laboratories and university programs offer unique resources for mounting aggressive programs to support the nation’s students and teachers in science, mathematics and engineering.”

Charles M. Vest et al.,
Task Force on the Future of Science Programs,
Critical Choices: Science, Energy, and Security,
October 13, 2003



The Broader Agenda: Global Discovery

We stand on the rim of a new era of global discovery. As science communication evolves due to the Internet, grid computing, simulation, collaboratories, and other technological advances, the opportunity exists to create a single-search interface for the whole of science.

Beyond the immediate focus on education, the DOE Science Accelerator will address major challenges in access to and use of science information—specifically, the huge gap in the Internet when it comes to accessing science databases. The DOE Science Accelerator will bridge that gap with robust and time-saving precision search tools.

The DOE Science Accelerator will advance R&D creativity through a world-class user facility for the testing and application of federated search, analysis and other emerging information science technologies. The DOE Science Accelerator will enable the information consumer—whether in high school, college, graduate or post-graduate school, or whether a researcher, teacher, policy maker, or simply a science-attentive citizen—to rapidly navigate through petabytes of information and, with precision, find the precious few nuggets required to advance the science project at hand.



The portability of knowledge is essential to workforce competitiveness.

The Office of Science's ability to demonstrate a high return on increased basic research funding depends, in large part, on the success with which scientific knowledge generated by its extensive R&D programs is diffused and re-used. Demonstrating this return on investment is increasingly important.

The DOE Science Accelerator specifically supports the Office of Science Strategic Plan Goal 7, Provide the Resource Foundations that Enable Great Science, and recommendations from national strategic efforts such as the American Competitiveness Initiative, the No Child Left Behind Act, the Academic Competitiveness Council, the National Commission on the Future of Higher Education, and the National Academies' "Rising Above the Gathering Storm" report.

7 Provide the Resource Foundations that Enable Great Science



OSTI's Unique Position

OSTI's unique STI collection

A repository invaluable to the science community:

- Over 1 million documents, classified and unclassified
- From the Manhattan Project to present, with daily additions
- Comprehensive and current
- Legacy research results not available anywhere else
- Over 125,000 full-text reports, fully searchable online
- Over 4 million R&D citations to research of interest to DOE

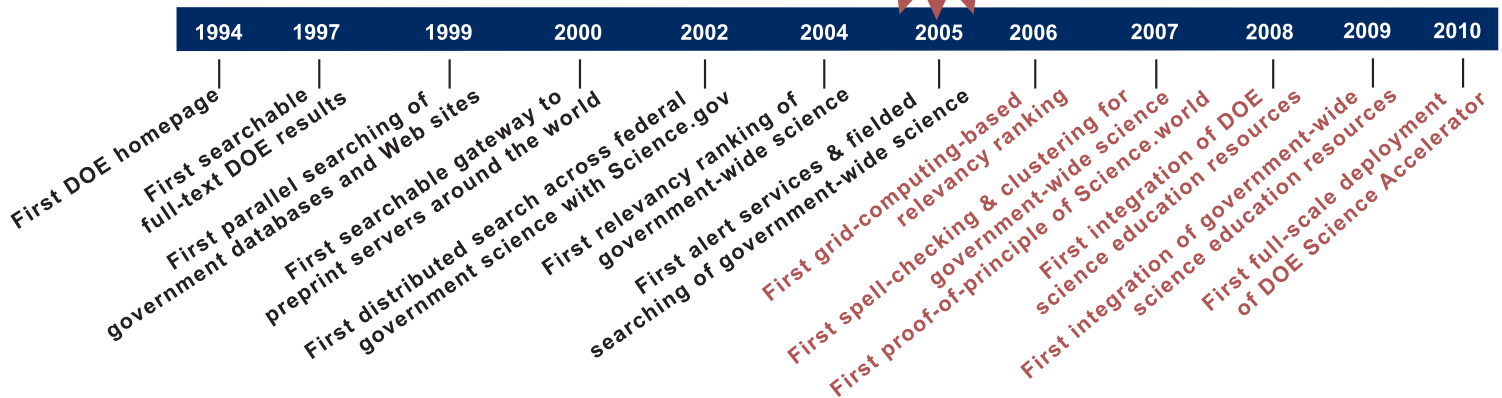
plus other R&D information, such as active research projects.

OSTI has a long history of recognizing gaps in scientific knowledge diffusion and developing innovative approaches for filling those gaps. In its early days, OSTI was charged with creating an agency-wide R&D information program from scratch. Over the next 50 years, OSTI capitalized on early adoption of technology, maximizing the use of computers for information search and retrieval. In the late 1990s, OSTI led the way among federal science agencies in going beyond electronic bibliographic information to bringing electronic full text to the desktops of scientists. To this day, the volume of DOE's electronic full-text R&D information far outpaces any other science agency.

As agencies began hosting Web-searchable R&D databases, the science community certainly benefited but was still left with an inefficient means to locate and navigate through disconnected, disparate collections. Again, OSTI recognized this and pioneered groundbreaking distributed precision searching technology with the launch of Science.gov. OSTI has brought Google-like capabilities to the deep Web through products such as E-print Network, Science Conferences and Federal R&D Project Summaries.

Now, OSTI sees another need: The huge gap in the Internet where search engines do not reach often obscure

40 million
knowledge
transactions
in FY '05

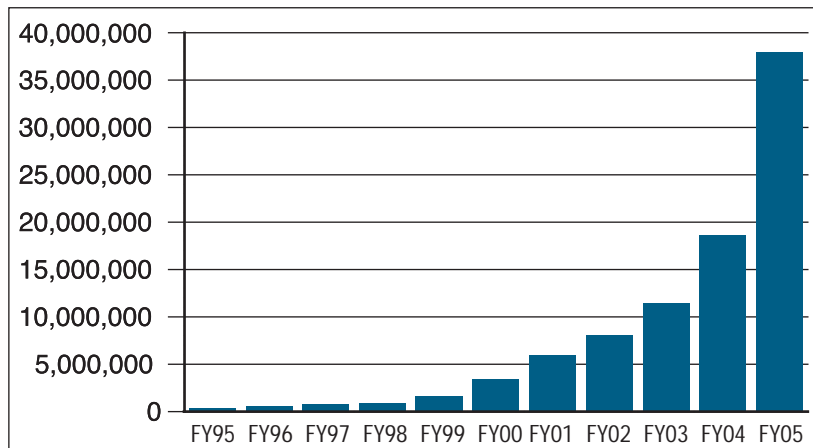


Milestones Pioneered by OSTI

Future Innovations

science databases, disconnected education resources and non-federal sources such as university and international R&D.

As in the past, OSTI also sees a solution: We must overcome the challenge of scalability to take distributed searching beyond its current capacity limitations and enhance precision searching and Web 2.0 applications, accommodating exponential increases in information, while still delivering results in seconds. We can do this because we have a history of rising to big challenges; we can do it because we have a technological path forward; and we can do it because we are energized by the role we will play in accelerating science.



OSTI Online Information Transactions

Since 1999, the number of annual transactions (page views and downloads) for science R&D information in OSTI's databases has increased from 1.5 million to 38 million—a 2,433 percent increase.

Statutory Authority

The Energy Policy Act of 2005 states: "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 activities supported by the Department."

The Atomic Energy Acts of 1946 and 1954, the Energy Reorganization Act of 1974, the Department of Energy Act of 1977, and the Energy Policy Act of 2005, all call for the dissemination of scientific and technical information (STI) to the public, especially information resulting from DOE and predecessor agency R&D.

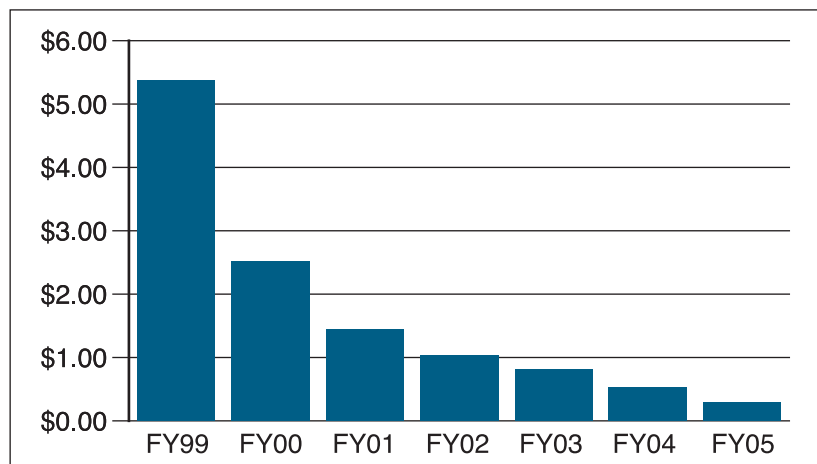
The Path Forward

The knowledge that will be created from increased R&D funding will only be useful if it is accessible; if it can be precision searched; if it is available across scientific communities; and if it has a facility to move easily between data and text. Accelerating knowledge diffusion accelerates the advancement of science.

Anticipated cost

The anticipated annual cost of the DOE Science Accelerator, starting in 2008, will be \$3 million. This cost covers several discrete components: Identifying science information sources; designing and developing sophisticated precision search algorithms; building distinct science education search and relevance ranking algorithms; Web-enabling other valuable science education materials; resolving language differences from global sources; integrating access between text and numeric data; and building innovative tools, such as clustering and visualization, to enhance the use and re-use of R&D results.

OSTI's deployment of groundbreaking information access technology has been developed only through the most judicious use of very limited funding. OSTI has demonstrated extreme efficiency in reducing its cost per transaction to the lowest among its federal counterparts. The National Institutes of Health (*NIH*), certainly a benchmark agency in research and the production of scientific literature, spends roughly 30 times more than DOE on getting its R&D message out. This investment contributes to the enviable level of public awareness of NIH's research programs.



OSTI Cost per Customer Transaction

Near-term objectives

- Implement distributed access to DOE's electronic educational resources
- Expand content to 100 additional sources
- Preserve and improve digital access to DOE's pre-1990 R&D literature
- Enhance precision searching
- Develop next-generation relevance ranking algorithms
- Enhance computing power to accommodate increased content and precision searching
- Develop prototype for analytical tools
- Develop prototype for grid computing
- Expand services to encourage collaboration and sharing of information, including RSS, alerts, blogs, tags, podcasts and other interactive technologies (*Web 2.0*)

Long-term goals

- Discover relevant data across the whole of science with a single entry point
- Overcome barriers to searching 1,000 databases in parallel
- Illuminate obscure databases
- Automate translation of queries and results
- Promote cross-discipline communication
- Integrate text and numeric data
- Speed diffusion of knowledge
- Accelerate scientific advancement

Conclusion

Just as science advances only if knowledge is shared, accelerating the sharing of knowledge will speed up the advancement of science. The DOE Science Accelerator will accelerate the sharing of knowledge by converting comprehensive cross-community searches from the impractical to the routine. Specific benefits include:

- Providing easily searched, relevant information, ranging from practical information for the consumer to highly technical scientific data for the research scientist
- Supporting future scientists and engineers with information and science education resources, initially converting DOE education resources from isolated islands of information to a virtual integrated whole
- Raising scientific and technical literacy

“The calculus of innovation is really quite simple:

- ***Knowledge drives innovation;***
- ***Innovation drives productivity;***
- ***Productivity drives our economic growth.***

That's all there is to it.”

William R. Brody
President
Johns Hopkins University
*U.S. Competitiveness:
The Innovation Challenge*
Testimony to the House
Committee on Science
July 21, 2005

