The Digital Road to Scientific Knowledge Diffusion

A Faster, Better Way to Scientific Progress?

By David E. Wojick, Walter L. Warnick, Bonnie C. Carroll, and June Crowe

Introduction

With the United States federal government spending over $130 billion annually for research and development, ways to increase the productivity of that research can have a significant return on investment. It is well known that all scientific advancement is based on work that has come before. Isaac Newton expressed this thought most eloquently in 1676, when he wrote, "If I have seen further than others, it is by standing on the shoulders of giants."

The process by which science knowledge is spread is called *diffusion*. It is therefore important to better understand and measure the benefits of this diffusion of knowledge. In particular, it is important to understand whether advances in Internet searching – such as simultaneous, ranked searching of distributed digital collections made broadly available via the Internet – can speed up the diffusion of scientific knowledge and accelerate scientific progress. Near-term opportunities continue to emerge to further speed up knowledge diffusion. To help craft a strategy for converting opportunities to reality, research is needed on the impact such speeding up of knowledge diffusion has on the advancement of science.

This article discusses these issues and describes research being conducted by the Office of Scientific and Technical Information (OSTI) of the United States Department of Energy (DOE) under its strategic initiative, Innovations in Scientific Knowledge and Advancement (ISKA).

Diffusion of Scientific Knowledge

Almost all communication, whether spoken or written, constitutes the sharing of knowledge. Although much of this knowledge is personal and local, our civilization is based on the widespread use of general knowledge. One of the most eloquent proponents of the diffusion of knowledge was Thomas Jefferson, who in 1786 said, "I think by far the most important bill in our whole code is that for diffusion of knowledge among the people. No other sure foundation can be devised for the preservation of freedom and happiness." [1]

Scientific publication, in all its myriad forms, is a huge system of deliberate knowledge diffusion. As collectors, organizers, and disseminators of published knowledge, all libraries, including digital libraries, have knowledge diffusion as their primary mission.

Science is organized into various research communities, each pursuing a particular set of scientific problems within their discipline. Most of the knowledge produced within a given community stays within that special scientific arena because the nexus of interactions – both interpersonal and through publication exchange – are there. However, as we see from the increasingly interdisciplinary nature of research today, research results from one community increasingly are useful in a wide array of other communities as well.

With the goal of advancing science, principal concerns are: knowing how knowledge flows *within* science, and knowing how new knowledge and technology flow *into* science. There are two different kinds of knowledge flow: (a) forward or within-community flow and (b) lateral or between-communities flow. These kinds of flow are of special interest, because diffusion of scientific knowledge is essential for scientific progress, and it is believed that a corollary exits such that as diffusion of knowledge is speeded up, scientific progress will be accelerated also.

Global Discovery

Increasingly science information is in digital format, and today, Internet search is the principal means by which the outward flow of this information is facilitated and determined. To address the complexity of the search issue, we use the term *global discovery* for the act of searching across heterogeneous environments and distant communities.

There are thousands of scientific research communities, with millions of researchers and thousands of journals and multiple data sources. It is challenging for a particular scientist to identify the appropriate combination of digital resources to expand his or her research horizons, but this challenge can be better met through global discovery. Global discovery has the potential to facilitate the advancement of science. If scientists could easily discover the initial breakthroughs being made in communities other than their own, then scientific knowledge diffusion would be greatly accelerated. Thus, global discovery itself has become a necessary focus area for research.

Significant strides have been made in global discovery in recent years, but the vast majority of scientific information resources continue to be held in deep web databases that many search engines cannot fully access. Some search engines such as Google Scholar are attempting to change this situation by harvesting small parts of the deep web, but at the time of writing, this remains an effort in progress.

Bibliographic databases offer limited information about resources; however, they only contain a tiny fraction of the total content of a resource, such as a 100-word abstract to a 10-page article or a 100-page report. The problem with this limited information is that it often leaves important content undiscoverable, because abstracts usually focus on community-specific aspects of the research while more global aspects of the research, such as mathematical technique, are barely touched on, if at all.

Thus, the basic problem is that while vast quantities of scientific information are available in principle via the web, there still is no simple way for a scientist to get to it all. The information is mostly available only in community-specific databases. Because of this distribution, in-depth global discovery still proceeds primarily on a community-by-community basis. In most cases, the labor involved in such searching has been prohibitive; thus, up till now global discovery has been achieved only to a very limited degree.

However, another way to improve global discovery is emerging, one that promises to greatly extend every scientist's capability. This new approach involves the simultaneous, ranked, federated, full text search of multiple, scientific databases across many communities. In principle, all the scientific content accessible via the web can be searched at once this way; in actuality, we are still far from that goal, because federated search is difficult and expensive to set up.

The OSTI Initiative to Improve Global Discovery

The mission of the Office of Scientific and Technical Information (OSTI) is to facilitate science by disseminating information, especially research results. Today OSTI includes a digital library with over 10 million pages online. There are over 100,000 DOE research reports available on-line, as well as several large bibliographic collections and numerous other resources.

OSTI has also been a leader in the simultaneous, ranked, searchable federation or aggregation of large digital collections that reside elsewhere. This includes creating the ePrint Network, which aggregates 35 science preprint databases, as well as operating Science.gov, the portal supported by a 14-agency consortium with over 50 million pages online. Currently, OSTI is engaged in a strategic initiative, Innovations in Scientific Knowledge and Advancement (ISKA), which focuses on these and related innovations.

As the ISKA initiative matures, its goal is to significantly enable and accelerate advances in science. To date, ISKA's program has been modest, but it is nevertheless worth noting. It includes the following key components:

1. Federation of distributed collections with simultaneous, ranked, full-text searching;
2. Distributed computer processing techniques to support such collections;
3. Integration of numeric data and textual information; and
4. Modeling scientific exchange in the research process.

Science.gov (http://www.science.gov) readily shows the value of investigating a topic by drawing information resources from a number of disparate scientific domains. Science.gov searches the open access literature, especially the so-called "gray literature" of research reports, preprints, etc. But this is still just a small fraction of the potentially available science information, and Science.gov is highly selective in some ways. OSTI also has other, smaller federations of digital resources and libraries with global discovery

capability. This interdisciplinary approach to scientific discovery is a small, yet important, blueprint of the global search facility of tomorrow. Specifically, what ISKA wants to show is that simultaneous, ranked search of distributed digital collections, made broadly available via the Internet, speeds up diffusion of scientific knowledge and accelerates scientific progress. In particular, it facilitates global discovery.

What We Know from Searching the Scientific Literature

Since the corollary to our premise that science depends on diffusion is that speeding up diffusion will accelerate scientific progress, it is important that we at ISKA have evidence to support the directions it is going and assertions being made. We undertook a literature search to see what research is being done and what is known about the diffusion of scientific knowledge, especially with regard to the assertion that speeding up knowledge diffusion will accelerate scientific progress. In particular, we were looking for insights, methods or tools that could be applied to measure or evaluate specific ways to speed up diffusion. Our findings were:

1. **Diffusion of scientific knowledge is not a major research area.** There appears to be a lot of scientific work on diffusion of knowledge as it relates to innovation and technology, but very little on diffusion of scientific knowledge within science and for the advancement of science, per se. Regarding the basic corollary, we found almost nothing on how – or if – speeding up the diffusion of scientific knowledge affects scientific progress. Although this may be taken as a given, finding empirical evidence is another matter. Based on our literature search, it appears that little or no scientific research has been done specifically on the basic corollary hypothesis that speeding up diffusion will accelerate scientific progress. The principal reason for this seems to be the lack of acceptable ways to measure the quantities in question, that is, the speed of diffusion and the acceleration of scientific progress. In the popular literature it is commonly said that the Internet will revolutionize science by increasing communication, but this assumption has not been translated into a scientific research program.
2. **A number of communities are actively engaged in scientific research on topics that are closely related to the diffusion of scientific knowledge.** These topics include the diffusion of innovation and research program evaluation. Likewise, there is ongoing research into the diffusion of knowledge in general, including digital information via the Internet. All of this research contributes to efforts to evaluate new technologies and diffusion strategies. However, the primary focus of most research is the diffusion of technology, or of knowledge in general, not science knowledge specifically. There appears to be no corresponding effort when it comes to speeding up diffusion of scientific knowledge and accelerating scientific progress. This lack of research may be a significant gap in our national research agenda. Particular emphasis should be given to the potential impact of new digital resources on the progress of science, including federation of collections to facilitate global discovery. These communities are summarized in Table 1.

Table 1. Related Research Communities on Knowledge Diffusion

Diffusion of innovation. A lot of work has been done on the diffusion of innovation, principally by economists, market researchers, and historians. However, innovation has been defined in most cases as technology in use, not scientific knowledge. Some quantitative work has been done, using measurable features of technology, especially statistics for manufacturing, sales, and usage. There is a heavy focus on new product development and marketing, as well as economic impact.

Knowledge management (KM) and diffusion. There seems to be very little scientific research in the area of knowledge management, perhaps because it is technology driven. Also KM is more about organizing knowledge than about diffusion.

Network theory. There is a large body of knowledge diffusion work based on information network theory, much of it focused on human interaction and communication. Internet communication is a major area of this network research. But this work is not directly related to scientific knowledge or scientific progress.

Semantic web. Semantic web research is focused on using semantic structures to integrate diverse databases. While this work may well be relevant to the development of tools to speed up the diffusion of scientific knowledge, it is not about such diffusion per se.

Science R&D program evaluation. The impact of specific science programs is an active area of research. Quantitative methods tend to focus on bibliometrics, especially citation analysis, including patent citations. However, most of the research is focused on the relation between funding and performance, measured by citation. While this does address the issue of diffusion of scientific knowledge, the concepts of speeding up diffusion and scientific progress are not well developed.

Industrial research policy. This is an active area of micro-economic research that tends to regard diffusion as a so-called "spillover" or free rider problem, where one firm benefits from another's expenditures. There is some interesting work on whether firms should be more open with their research results so that more diffusion occurs, along the lines of federal and university research. This might be considered as a special case of the hypothesis that speeding up diffusion accelerates progress.

Patent analysis. The number of patents has in some cases been used as a surrogate for knowledge. In this case there is an active debate about whether the increase in annual U.S. patenting rates in recent decades is in fact an acceleration in progress. But again, a focus on patents is basically a focus on technology, not scientific knowledge or progress.

Philosophy of science and science studies. There is some work on the influence of social factors on the progress of science, including the role of scientific communities. There is also work on the application of mathematical logic to the structure of scientific

revolutions.

3. **Conceptual complexity is an obstacle to broad scientific search.** Regarding our search itself, we found that the conceptual complexity of knowledge diffusion makes this kind of broad search difficult. Important aspects of diffusion are the subject of research by a number of distinct communities, using a variety of different concepts and vocabularies. Moreover, many key concepts, including "diffusion" and "knowledge," are not well defined.

    - It is likely that this sort of conceptual complexity occurs frequently when one looks across many diverse research communities for a common underlying subject. This result has implications for the issue of diffusion of scientific knowledge in general and global discovery in particular. Given that semantic tools and semantic web research are looking at the issue of multiple technical languages, it seems likely that they will eventually have something to offer in overcoming this obstacle to coherent search. But one recognizes that semantics research is in the early stages of development and to date does not appear to have been applied to the specific problem discussed in this article.

4. **A coherent program of research into the diffusion of scientific knowledge and its relation to scientific progress does not seem to exist at this time.** Research is needed on how to speed up diffusion of scientific knowledge and looking at specific approaches, such as simultaneous, ranked search, federated digital means, and digital libraries to do so, as well as what impact this speeding up might have on scientific progress.

5. **Semantic tools are needed to facilitate full text search of scientific information, as a means for speeding up diffusion.** Research communities are also language communities, each with its own technical vocabulary. Very different languages may be used to talk about what are in fact closely related concepts. This is a significant obstacle to global search and discovery, one that semantic tools may help resolve.

Conclusion

There are three strategies to better understand and promote the advancement of science:

1. Developing the conceptual framework to understand scientific knowledge diffusion and clarification of concepts so that the diffusion of scientific knowledge corollary can be proven;
2. Investigating the body of knowledge that exists, identifying gaps, and defining areas of research; and
3. Promoting the development of tools for global discovery that should be tested for the impact on the advancement of search and enabling of global discovery.

Through all three strategies, the advancement of the science corollary is addressed.

The OSTI ISKA initiative focuses on the third strategy, but in the process of setting out its agenda has confronted both the need for conceptual context and the importance of reviewing the body of knowledge that exists. With $130 billion being spent annually on federal science and technology, this is a significant enough challenge that it should be more comprehensively addressed. The payoff is that the greater sharing of intellectual endeavors will enhance the discovery process and thereby enable scientific breakthroughs to occur more quickly, for the ultimate benefit of society.

Reference

[1] Rayner, B.L.; revised and edited by Coates, Eyler Robert, *Life of Thomas Jefferson*, <http://etext.virginia.edu/jefferson/biog/lj13.htm> (accessed 17 January 2006).

David E. Wojick, Consultant, Innovations in Scientific Knowledge and Advancement, DOE Office of Scientific and Technical Information; Walter L. Warnick, Director, DOE Office of Scientific and Technical Information; Bonnie C. Carroll, President, Information International Associates, Inc.; and June Crowe, Senior Researcher, Information International Associates, Inc.