12-3-2010

Network Science: Using Information Technology to Enable Collaboration

Arden L. Bement Jr.
Purdue University, bement@purdue.edu

Talk at the University of Southern California
"Collaboration in the Academy" Retreat
Norman Lear Center
Annenberg School for Communications
December 3, 2010

"Networked Science: Using Information Technology to Enable Collaborative Research"
Arden L. Bement, Jr.

Good afternoon! I am greatly honored to be invited to give remarks at this retreat on the subject "Networked Science: Using Information Technology to Enable Collaborative Research". I am coming at this subject as a science administrator, not as a practitioner. Therefore, much of what I have to say will be a reprise of the rich discussion this morning.

I am delighted to be back on the campus of USC, which is prominent in my memories. During my tenure at TRW during the 1980s, I became aware of the special relationships between TRW and USC through the outstanding service on the Board of Trustees of the late Virginia Ramo, the evening courses taught by USC faculty at TRW's Space Park, and my close working relationship with Leonard M. Silverman, former Dean of the Viterbi School of Engineering. Over the years my interest in USC has been sustained by the great leadership of Steven Sample in taking USC to a higher level of excellence and the appointment of Si Ramo as a Presidential Chair and Professor of Electrical Engineering. Si was a great role model and inspiration for all scientists and engineers at TRW. He taught us that lifetime learning was not only an obligation but also a privilege.

During the information revolution, which continues to build momentum, there are a number of changes in the international scene that have driven enormous changes in how science is being done and how scientists engage with other scientists in performing science. First of all the tempo has increased due to increases in scientific productivity. This is driven in part by international competition. Most economists in the world will agree that economic growth in today's information revolution is substantially linked to investments in education, research, and infrastructure, especially cyberinfrastructure. This has naturally led to heightened investments in these areas, resulting in a greater worldwide leveling in research intensity relative to GDP. In addition to increased competition in the conduct of science there is a greater imperative among governments to measure the returns from scientific discovery in terms of new job creation and innovative activity.

There has also been a movement among U.S. universities to become international universities through heightened international enrollment, student exchanges, research collaborations and the establishment of campuses abroad. Purdue University is an example. It ranks among the top four universities in international enrollment (USC ranking first), and has prominently emphasized global engagement and leadership in its strategic planning. The establishment of the Global Policy Research Institute, which I have the honor to lead, is one example among many of Purdue's global leadership.

International networking has expanded substantially in just the past five years. While there are still some dark spots, most of the major academic, industrial, and government research centers in the world are now connected. I expect that even the dark spots will soon be connected.

Because of these developments the world continues to shrink at an accelerating rate. There is a growing imperative to collaborate in order to compete. To do frontier research it is first necessary for the scientist to know where the frontier is, where the next salient at the frontier is likely to occur, and what grand challenge questions need to be addressed. Because of the increased intensity of global research, insights into finding answers to these questions might be quite different in different parts of the world. Hence, it is important for scientists to network globally to keep from becoming "blind sided". To guard against this, many scientists are now connecting more extensively with virtual international research groups. While such global outreach among scientists is not new, the rapid growth of cyber infrastructure, the accelerating volumes of data and information flows, and the complexity of modern-day scientific problems make global engagement today more "sporty".

Among the factors at play is the growing attention being given to interdisciplinary research involving research groups and centers, computer modeling and simulation, and data sharing and synthesis. I believe that these trends are driven by faculty interests in doing research systemically to tackle the complexities inherent in global, grand-challenge problems by using the advanced computational tools at their disposal.

Finally, pressures for high-end research instrumentation and facilities with embedded cyber tools and networks continue to skyrocket. NSF is under increasing pressure to increase its cap on major research instrumentation to as much as $100 million. Furthermore, the pressures to build new national research facilities costing from $500 million to above $1 billion are likewise increasing, not just in the U.S. but also around the world. This trend raises the question "Should we not be building more kinds of facilities rather than more facilities of one kind?" It was the pondering of this question that led to the development and commissioning of the National Earthquake Engineering System (NEES) connected by a common backbone network and now managed by Purdue. USC is one of the 14 NEES system members.

The NEES network has been a resounding success not only in providing the means for free, on-line data sharing but also enabling coupled system testing of virtually-connected structures. Having demonstrated the benefits of such a virtual test network at the national scale it was natural to propose that such a network could and should be extended internationally among seismically-active nations. I was delighted in 2005 to sign an MOU between the NSF and MEXT* to connect Japan's E-Defense shake table, the largest and most complex test facility of its type, to the NEES network. Hopefully, other countries will join the NEES network in the near future.

One can cite other examples of virtual networks. Examples would include the autonomous, virtual operational scheduling and sharing of data from major telescopes, such as the Gemini telescopes in Chile and Hawaii and the 10-meter telescope at the South Pole. Observational "virtual" scientific networks funded by the NSF that are connected by a common network and data repository include the Incorporated Research Institutes for Seismology (IRIS), the Earth Observing Network (Earthscope), Ocean Observing Initiative (OOI), National Ecological Observing Network (NEON), and Arctic Observing Network (AON). Such virtual facilities will be major contributors to the burgeoning volume of data and information flows that are already saturating our networks, storage facilities and capabilities to curate and evaluate data.

In the remainder of my remarks I would like to outline in the way of examples some of Purdue's

*Japan's Ministry of Education, Culture, Sports, Science and Technology contributions to facilitating virtual communities through the open sharing of data, metadata, and publications. It should be stressed at the outset that problems in open sharing among these three categories are not just problems of scale but also problems of scope concerning scientists, such as trust, reliability, security, provenance, precedence, and authenticity.

The growth in high-performance computing occurring not only at Purdue University but also throughout the U.S. is both driving and responding to the rapid acceleration in data generation. Purdue employs community clusters, which tap the operational availability of campus computers broadly in the interest of reducing cost and power consumption while improving versatility. Since 2006, computing power at Purdue has increased more than ten times. But since 2000 it has increased 200 times, or two orders of magnitude in one decade. To meet this demand Purdue built in 2008, 2009, and 2010 three supercomputers with more than 130 teraflops of collective capacity. These three supercomputers are named the “Steele”, “Coates”, and “Rossmann”, respectively, after three campus pioneers in the use of computers. The Coates and Rossmann supercomputers were ranked, respectively, at the 2010 supercomputing conference the 147^{th} and 126^{th} among the world’s 500 most powerful known computing systems.

In addition to these three notable supercomputers, Purdue has developed a TeraGrid resource called “Wispy”, a cloud computing system, which allows researchers to package self-contained programs, or “virtual machines”, to run remotely on a specifically configured cluster supercomputer. Up to 128 virtual machines can run on a Wispy cluster at a time and address a variety of computing environments. Since a Wispy cluster is readily scalable, this capability can greatly extend computational capacity over time at low cost.

Perhaps the innovation Purdue is most noted for in support of virtual communities is the development of HUBzero, which is the open-source foundation for the development of a growing number of HUBs, one of the most notable being the nanoHUB. HUBzero has been termed a Facebook, “super-charged with steroids”, for scientists. For those of you who may be unfamiliar with a HUB, it is a web-based collaboration environment with such features as how-to videos, interactive simulation tools, online presentations, sites for upgrading new resources, tool development, ratings, citations and content tagging. It is a sharing, learning, exploring, discovering, and virtual people -connecting environment. For example, the NEEShub has been developed to access NEES projects, run earthquake simulators, facilitate the analysis of structural responses to displacement reversals, support learning with data and simulations, and share research results and publications.

Open, “free to access” data and information flows are essential for scientists in general and virtual communities in particular. Data mining of astronomy and astrophysics data sources, for example, such as the National Virtual Observatory, NASA’s Astrophysics Data System, and international astronomical and astrophysics surveys, such as the Sloan digital sky survey, have enabled even young amateur astronomers to discover exoplanets, galactic degeneracies, and binary stars.

Scientists who made discoveries by data synthesis rather than by original research were not highly regarded in the past. However, this has changed with open data and information sharing and the availability of computational tools that enable data mining, synthesizing, modeling, and simulating, and thereby the discovery of new concepts.

Data storage, archiving, safeguarding, and making accessible data, papers, and publications have brought university libraries even closer to scientists as full partners in their discovery and publication endeavors. Supercomputers and object-based storage systems are now common tools for libraries.

Furthermore libraries are now much more involved in merging semantics, ontology and data base development. University libraries collaborate actively in establishing innovative means to perform these roles. For example, many new "light archives" have been established to circulate e-publications and digitized archives broadly as compared with "dark archives" where they may never see the light of day.

Purdue libraries are especially active in performing these many roles. Among their active projects are the following:

- Archiving, safeguarding and making accessible to Indiana state residents U.S. federal documents.
- Establishing an Entrepreneurship Business Information Network (**e-Bin**) aimed at reaching out to entrepreneurs in 14 countries during its pilot phase.
- Building a Distributed Data Curation Center (**D2C2**) to help scientists determine how to best archive and preserve their data. This center, facilitated by 32 terabytes of data storage provided by Sun Microsystems, being developed in partnership with the University of Illinois at Urbana-Champaign, will provide data curation profiles across multiple research disciplines. It will address the questions: "Which researchers are willing to share data, when, with whom, and under what circumstances?"
- Building "**Purdue e-Scholar**", which is a distributed, institutional repository containing collections of digital documents, such as doctoral dissertations, faculty-authored journal articles, conference proceedings, and technical reports (provided by "**Purdue e-Pubs"**); digitized archival collections (provided by **Purdue e-Archives)**; and research datasets (provided by **Purdue e-Data**)
- Using the content of the Open Access Repositories (**OpenDOAR**) to harvest, characterize, test and measure a randomly selected set of 100 metadata records to establish criteria for establishing collections of such records.
- Engaging with the CIC-Google book digitization project to scan as many as 500,000 to 750,000 volumes from the Purdue Library's signature collection. This partnership marks the first time that a consortium of independent institutions (CIC) has joined in a collective scanning of print materials (Google) with the potential of digitizing up to 10 million volumes.

Most of the developments outlined above in computation, HUB development, and open data and information sharing also benefit education. The open courseware movement, initiated in 1999 by the University of Tübingen, was soon followed in 2002 with the launch of MIT OpenCourseWare, and subsequently joined by several other universities. This movement has provided enormous benefits to science, technology, engineering, and mathematics (STEM) education internationally, especially for developing countries and sparsely populated regions of the world. Advanced information and communication technologies are now commonly deployed in learning laboratories at most major U.S. universities.

Looking to the future, one can expect additional revolutionary developments in how research and STEM education are conducted. For example, the evolution of a semantic web, the dream of Tim Berners-Lee to make available machine-readable metadata that can facilitate communications among machines, will be advanced by new scientific understandings of how the human brain stores and transfers data.

The size of the neuroscience community has increased dramatically in recent years due to interdisciplinary couplings with other disciplines. The potential for increasing exchanges of insights should bring about major advances in understanding the functionality of the human brain in greater scientific detail. Such advances in knowledge should advance prospects of developing “natural” bridges between human intelligence and machine “intelligence”, enabling greater synergies between what the human can do best and what the machine can do best. This will in turn lead to new advances in how scientists do science and how virtual communities of scientists collaborate. These advances will also lead to further improvements in the efficiency of manufacturing and operating systems, such as a “smart grid”, through the use of both central and networked distributed controllers that have embedded anticipatory and “smart agent” functions.

A Blue Ribbon Advisory Panel of the National Science Foundation in their 2003 report “Revolutionizing Science and Engineering Through Cyberinfrastructure” has expressed this vision:

“…a new age has dawned in scientific and engineering research, pushed by continuing progress in computing, information, and communication technology and pulled by the expanding complexity, scope, and scale of today’s challenges…The emerging vision is to use cyberinfrastructure to build more ubiquitous, comprehensive digital environments…that can serve individuals, teams, and organizations in ways that revolutionize what they do, how they do it, and who participates.

This vision provides hope that as the world continues to shrink, and virtual engagement is further facilitated, scientists collaboratively will more readily find solutions to the daunting global challenges facing the world today and pave the way to an improved quality of life for humanity.

Thank you!