An Endless Frontier Postponed

Next month, U.S. scientists Vinton G. Cerf and Robert E. Kahn will receive computing's highest prize, the A. M. Turing Award, from the Association for Computing Machinery. Their Transmission Control Protocol (TCP), created in 1973, became the language of the Internet. Twenty years later, the Mosaic Web browser gave the Internet its public face. TCP and Mosaic illustrate the nature of computer science research, combining a quest for fundamental understanding with considerations of use. They also illustrate the essential role of government-sponsored university-based research in producing the ideas and people that drive innovation in information technology (IT).

Recent changes in the U.S. funding landscape have put this innovation pipeline at risk. The Defense Advanced Research Projects Agency (DARPA) funded TCP. The shock of the Soviet satellite Sputnik in 1957 led to the creation of the agency, which was charged with preventing future technological surprises. From its inception, DARPA funded long-term nonclassified IT research in academia, even during several wars, to leverage all the best minds. Much of this research was dual-use, with the results ultimately advancing military systems and spurring the IT industry.

U.S. IT research grew largely under DARPA and the National Science Foundation (NSF). NSF relied on peer review, whereas DARPA bet on vision and reputation, complementary approaches that served the nation well. Over the past 4 decades, the resulting research has laid the foundation for the modern microprocessor, the Internet, the graphical user interface, and single-user workstations. It has also launched new fields such as computational science. Virtually every aspect of IT that we rely on today bears the stamp of federally sponsored research. A 2003 National Academies study provided 19 examples where such work ultimately led to billion-dollar industries, an economic benefit that reaffirms science advisor Vannevar Bush's 1945 vision in *Science: The Endless Frontier*.

However, in the past 3 years, DARPA funding for IT research at universities has dropped by nearly half. Policy changes at the agency, including increased classification of research programs, increased restrictions on the participation of noncitizens, and "go/no-go" reviews applied to research at 12- to 18-month intervals, discourage participation by university researchers and signal a shift from pushing the leading edge to "bridging the gap" between fundamental research and deployable technologies. In essence, NSF is now relied on to support the long-term research needed to advance the IT field.

Other agencies have not stepped in. The Defense Science Board noted in a recent look at microchip research at the Department of Defense (DOD): "[DARPA's] withdrawal has created a vacuum . . . The problem, for DOD, the IT industry, and the nation as a whole, is that no effective leadership structure has been substituted." The Department of Homeland Security, according to a recent report from the President's Information Technology Advisory Committee, spends less than 2% of its Science and Technology budget on cybersecurity, and only a small fraction of that on research. NASA is downsizing computational science, and IT research budgets at the Department of Energy and the National Institutes of Health are slated for cuts in the president's fiscal year 2006 budget.

These changes, combined with the growth of the discipline, have placed a significant burden on NSF, which is now showing the strain. Last year, NSF supported 86% of federal obligations for fundamental research in IT at academic institutions. The funding rate for competitive awards in the IT sector fell to 16%, the lowest of any directorate. Such low success rates are harmful to the discipline and, ultimately, to the nation.*

At a time when global competitors are gaining the capacity and commitment to challenge U.S. high-tech leadership, this changed landscape threatens to derail the extraordinarily productive interplay of academia, government, and industry in IT. Given the importance of IT in enabling the new economy and in opening new areas of scientific discovery, we simply cannot afford to cede leadership. Where will the next generation of groundbreaking innovations in IT arise? Where will the Turing Awardees 30 years hence reside? Given current trends, the answers to both questions will likely be, "not in the United States."

Edward D. Lazowska and David A. Patterson

Edward D. Lazowska holds the Bill & Melinda Gates Chair in Computer Science & Engineering at the University of Washington. David A. Patterson holds the E. H. and M. E. Pardee Chair of Computer Science at the University of California, Berkeley, and is president of the Association for Computing Machinery. Both are members of the National Academy of Engineering and the President's Information Technology Advisory Committee, and past chairs of the Computing Research Association.

*The House Science Committee will consider these issues at a 12 May hearing on "The Future of Computer Science Research in the U.S." See http://www.cra.org/research.

10.1126/science.1113963

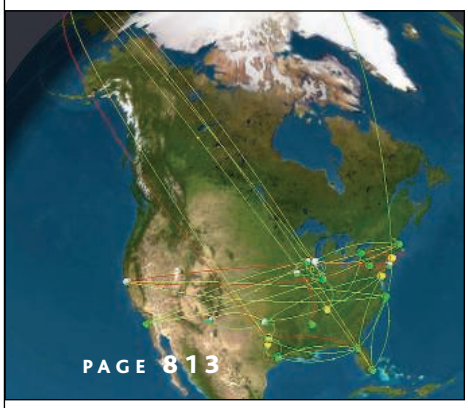
INTRODUCTION

All for One and One for All

As scientific instruments become ever more powerful, from orbiting observatories to genome-sequencing machines, they are making their fields data-rich but analysis-poor. Ground-based telescopes in digital sky surveys are currently pouring several hundred terabytes (10^{12} bytes) of data per year into dozens of archives, enough to keep astronomers busy for decades. The four satellites of NASA's Earth Observing System currently beam down 1000 terabytes annually, far more than earth scientists can hope to calibrate and analyze. And looming on the horizon is the Large Hadron Collider, the world's largest physics experiment, now under construction at CERN, Europe's particle physics lab near Geneva. Soon after it comes online in 2007, each of the five detectors will be spewing out several petabytes (10^{15} bytes) of data—about a million DVDs' worth—every year.

These and similar outpourings of information are overwhelming the available computing power. Few researchers have access to the powerful supercomputers that could make inroads into such vast data sets, so they are trying to be more creative. Some are parceling big computing jobs into small work packages and distributing them to underused computers on the Internet. With this strategy, insurmountable tasks may soon become manageable.

One approach to such "distributed computing" was pioneered by computer scientists working with SETI, the Search for Extraterrestrial Intelligence. The phenomenally successful SETI@home program now makes use of the idle computer time of millions of ordinary computer users, working as a screen saver to quietly crunch away at radio-signal data from deep space. As John Bohannon describes on p. 810, the same screensaver technique is now being used by a wide array of researchers studying everything from climate change to gravitational waves and protein folding. Bohannon also delves into the strange tribal world (p. 812) of the "crunchers": computer enthusiasts whose goal is to become the most prolific processors of data for various screen-saver research projects. And on p. 813, Mark Buchanan samples a piece of computer navel gazing: a distributed computing project to study the geography of the Internet itself.


PAGE 813

Another way of distributing both data and computing power, known as grid computing, taps the Internet to put petabyte processing on every researcher's desktop. On p. 814, Foster highlights the development of a lingua franca of grid computing: a set of standardized interfaces and protocols that permits researchers to work across the Web. Hey and Trefethen (p. 818) describe the U.K.-based e-Science program to design plug-and-play grid technologies for a range of disciplines. And Buetow (p. 822) outlines the ways in which cyberinfrastructure can weld together the vastly different styles of biomedical research.

For all the excitement, however, there are disturbing trends in the directions being taken by funding agencies that have historically been involved with driving the Internet revolution. In their Editorial (p. 757), Lazowska and Patterson consider how downsizing and short-term thinking threaten to derail the next generation of information innovation.

–DANIEL CLERY AND DAVID VOSS

CONTENTS

See also the Editorial on p. 757, News of the Week story by Daniel Clery, and related STKE material on p. 751 and at www.sciencemag.org/sciext/computers.

CREDITS: IOSIF LEGRAND/CALTECH

Science

Published by AAAS

DISTRIBUTED COMPUTING

NEWS

Grassroots Supercomputing

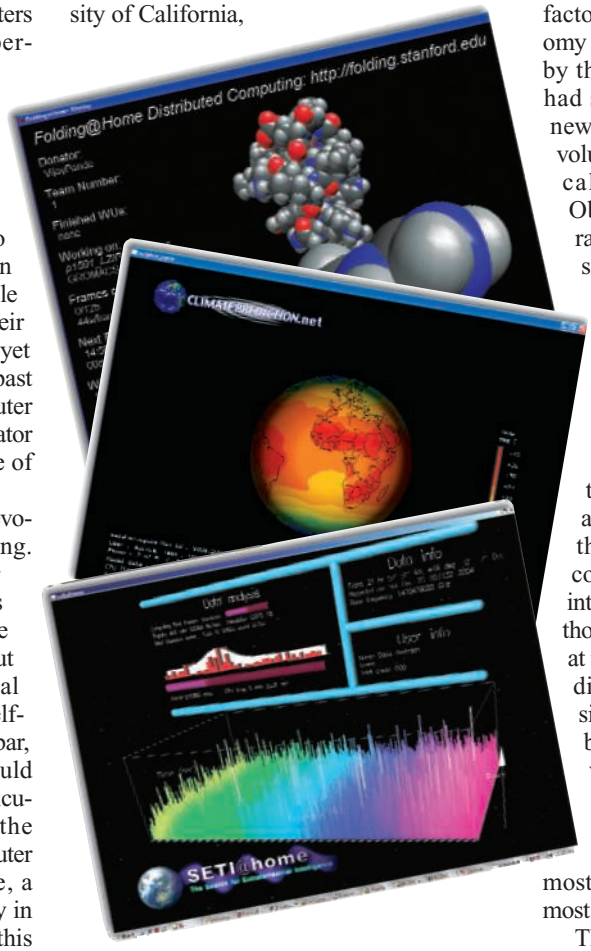
What started out as a way for SETI to plow through its piles of radio-signal data from deep space has turned into a powerful research tool as computer users across the globe donate their screen-saver time to projects as diverse as climate-change prediction, gravitational-wave searches, and protein folding

OXFORD, U.K.—If Myles Allen and David Stainforth had asked for a supercomputer to test their ideas about climate change, they would have been laughed at. In order to push the limits of currently accepted climate models, they wanted to simulate 45 years of global climate while tweaking 21 parameters at once. It would have required a supercomputer's fully dedicated attention over years, preempting the jealously guarded time slots doled out to many other projects. "Doing this kind of experiment wasn't even being considered," recalls Stainforth, a computer scientist here at Oxford University. So instead, he and Oxford statistician Allen turned to the Internet, where 100,000 people from 150 countries donated the use of their own computers—for free. Although not yet as flexible, their combined effort over the past 2 years created the equivalent of a computer about twice as powerful as the Earth Simulator supercomputer in Yokohama, Japan, one of the world's fastest.

Stainforth's project is part of a quiet revolution under way in scientific computing. With data sets and models growing ever larger and more complex, supercomputers are looking less super. But since the late 1990s, researchers have been reaching out to the public to help them tackle colossal computing problems. And through the selfless interest of millions of people (see sidebar, p. 812), it's working. "There simply would not be any other way to perform these calculations, even if we were given all of the National Science Foundation's supercomputer centers combined," says Vijay Pande, a chemical biologist at Stanford University in Palo Alto, California. The first fruits of this revolution are just starting to appear.

World supercomputer

Strangely enough, the mass participation of the public in scientific computing began with a project that some scientists believe will never achieve its goal. In 1994, inspired by the 25th anniversary of the moon landing, software designer David Gedye wondered "whether we would ever again see such a singular and positive event," in which people across the world join in wonder. Perhaps the only thing that could have that effect, thought Gedye, now based in Seattle, Washington, would be the discovery of extraterrestrial intelligence. And after teaming up with David Anderson, his former computer science professor at the University of California,



Strength in numbers. Millions of computers now crunch data for diverse research projects.

Berkeley, and Woody Sullivan, a science historian at the University of Washington, Seattle, he had an idea how to work toward such an event: Call on the public to get involved with the ongoing Search for Extraterrestrial Intelligence (SETI) project.

In a nutshell, SETI enthusiasts argue that we have nothing to lose and everything to gain by scanning electromagnetic radiation such as radio waves—the most efficient method of interstellar communication we know of—from around the galaxy to see if anyone out there is broadcasting. After the idea for SETI was born in 1959, the limiting factor at first was convincing radio astronomy observatories to donate their help. But by the mid-1990s, several SETI projects had secured observing time, heralding a new problem: how to deal with the huge volume of data. One Berkeley SETI project, called SERENDIP, uses the Arecibo Observatory in Puerto Rico, the largest radio telescope in the world, to passively scan the sky around the clock, listening to 168 million radio frequencies at once. Analyzing this data would require full-time use of the Yokohama Earth Simulator, working at its top speed of 35 teraFLOPS (10^{12} calculations per second).

Gedye and his friends approached the director of SERENDIP, Berkeley astronomer Daniel Werthimer, and posed this idea: Instead of using one supercomputer, why not break the problem down into millions of small tasks and then solve those on a million small computers running at the same time? This approach, known as distributed computing, had been around since the early 1980s, but most efforts had been limited to a few hundred machines within a single university. Why not expand this to include the millions of personal computers (PCs) connected to the Internet? The average PC spends most of its time idle, and even when in use most of its computing power goes untapped.

The idea of exploiting spare capacity on PCs was not a new one. Fueled by friendly competition among hackers, as well as cash prizes from a computer security company, thousands of people were already using their PCs to help solve mathematical problems. A trailblazer among these efforts was GIMPS, the Great Internet Mersenne Prime Search, named after the 16th century French monk who discovered a special class of enormous numbers that take the form $2^P - 1$ (where P is a prime). GIMPS founder George Woltman, a programmer in Florida, and Scott Kurowski,

CREDITS (TOP TO BOTTOM): PHOTOS.COM; FOLDING@HOME; CLIMATEPREDICTION.NET; SETI@HOME

Published by AAAS

a programmer in California, automated the process and put a freely downloadable program on the Internet. The program allowed PCs to receive a task from the GIMPS server, "crunch" on it in the background, and send the results back without the PC user even noticing.

Using computer time in this way is not always a blameless activity. In 1999, system administrator David McOwen marshaled hundreds of computers at DeKalb Technical College in Clarkston, Georgia, to crunch prime numbers with a program from a distributed network—but without getting permission. When found out, he was arrested and accused of costing the college more than $400,000 in lost bandwidth time. But the case never came to court, and McOwen accepted penalties of 80 hours of community service and a $2100 fine. The previous year, computer consultant Aaron Blosser got the computers of an entire Colorado phone company busy with GIMPS. Because his supervisor had given him permission to do so, he was not charged, but because at the time it was considered a potential act of Internet terrorism, the FBI confiscated his computers.

Undaunted, Gedye and his team set about carving up the SETI processing work into bite-sized chunks, and in 1999 the team went public with a screen-saver program called SETI@home. As soon as a PC went idle, the program went to work on 100-second segments of Arecibo radio data automatically downloaded from the Internet, while the screen saver showed images of the signal analysis. It took off like wildfire. Within 1 month, SETI@home was running on 200,000 PCs. By 2001, it had spread to 1 million. Public-resource computing, as Anderson calls it, was born.

So far at least, SETI@home hasn't found an ET signal, admits Anderson, and the portion of the galaxy searched "is very, very limited." But the project has already accomplished a great deal: It not only fired up the public imagination, but it also inspired scientists in other fields to turn to the public for help tackling their own computing superproblems.

Democratizing science?

Stanford's Pande, who models how proteins fold, was among the first scientists to ride the public-resource computing wave. Proteins are like self-assembling puzzles for which we know all the pieces (the sequence of amino acids in the protein backbone) as well as the final picture (their shape when fully folded), but not what happens in between. It only takes microseconds for a typical protein to fold itself up, but figuring out how it does it is a computing nightmare. Simulating nanosecond slices of folding for a medium-sized protein requires an entire day of calculation on the fastest machines and years to finish the job. Breaking through what Pande calls "the microsecond barrier" would not only help us understand the physical chemistry of normal proteins, but it could also shed light on the many diseases caused by misfolding, such as Alzheimer's, Parkinson's, and Creutzfeldt-Jakob disease.

A year after SETI@home's debut, Pande's research group released a program called Folding@home. After developing new methods to break the problem down into workable chunks, they crossed their fingers, hoping that enough people would take part. For statistical robustness, identical models with slightly tweaked parameters were doled out in parallel to several different PCs at once, so success hinged on mass participation.

The simulations flooded back. By the end of its first year, Folding@home had run on 20,000 PCs, the equivalent of 5 million days of calculation. And the effort soon proved its worth. Pande's group used Folding@home to simulate how BBA5, a small protein, would fold into shape starting only from the protein's sequence and the laws of physics. A team led by Martin Gruebele, a biochemist at the University of Illinois, Urbana-Champaign, tested it by comparing with real BBA5. The results, reported in 2002 in *Nature*, showed that Folding@home got it right. This marks "the first time such a convergence between theory and experiment could be made," says Pande.

Public-resource computing now has the feel of a gold rush, with scientists of every stripe prospecting for the bonanza of idle computing time (see table, left). Biological projects dominate so far, with some offering screen savers to help study diseases from AIDS to cancer, or predict the distribution of species on Earth. But other fields are staking their own claims. Three observatories in the United States and Germany trying to detect the fleeting gravitational waves from cataclysmic events in space—a prediction of Einstein's—are doling out their data for public crunching through a screen saver called Einstein@home. Meanwhile, CERN, the European particle physics laboratory near Geneva, Switzerland, is tapping the public to help design a new particle accelerator, the Large Hadron Collider. LHC@home simulates the paths of particles whipping through its bowels.

Project/URL	Research Base	Goal
Mersenne Prime Search www.mersenne.org	Worldwide	Identify enormous prime numbers
SETI@home setiathome.ssl.berkeley.edu	UC Berkeley	Find extraterrestrial intelligence
Folding@home folding.stanford.edu	Stanford	Predict how proteins fold
ClimatePrediction.net climateprediction.net	Oxford	Test models of climate change
LHC@home lhcathome.cern.ch	CERN	Model particle orbits in accelerator
Einstein@home einstein.phys.uwm.edu	U.S. and Germany	Identify gravitational waves
Cancer Research Project www.grid.org/projects/cancer	NCI and Oxford	Search for candidate drugs against cancer
Lifemapper www.lifemapper.org	University of Kansas	Map global distribution of species

The projects launched so far have only scraped the surface of available capacity: Less than 1% of the roughly 300 million idle PCs connected to the Internet have been tapped. But there are limits to public-resource computing that make it impractical for some research. For a project to make good use of the free computing, says Stainforth, "it has to be sexy and crunchable." The first factor is important for attracting PC owners and persuading them to participate. But the second factor is "absolutely limiting," he says, because not all computational problems can be broken down into small tasks for thousands of independent PCs. "We may have been lucky to have chosen a model that can be run on a typical PC at all," Stainforth adds.

In spite of those limitations, the size and number of public-resource computing projects is growing rapidly. Much of this is thanks to software that Anderson developed and released last year, called Berkeley Open Infrastructure for Network Computing (BOINC). Rather than spending time and money developing their own software, researchers can now use BOINC as a universal template for handling the flow of data. In a single stroke, says Anderson, "this has slashed the cost of creating a public-resource computing project from several hundreds of thousands of dollars to a few tens of thousands." Plus, BOINC vastly improves the efficiency of the entire community by allowing PCs to serve several research projects at once: When one project needs a breather, another can swoop in rather than leaving the PC idle.

Published by AAAS

It works, too

As the data streams in from the many projects running simultaneously on this virtual supercomputer, some researchers are getting surprising results. To the initial dismay of CERN researchers, LHC@home occasionally produced very different outputs for the same model, depending on what kind of PC it ran on. But they soon discovered that it was caused by "an unexpected mathematical problem," says François Grey, a physicist at CERN: the lack of international standards for handling rounding errors in functions such as exponential and tangent. Although the differences between PCs were minuscule, they were amplified by the sensitive models of chaotic particle orbits. The glitch was fixed by incorporating new standards for such functions into the program.

The results of ClimatePrediction.net have been surprising for a different reason. "No one has found fault with the way our simulations were done," says Stainforth. Instead, climate scientists are shocked by the predictions. Reporting last January in *Nature*, a team led by Stainforth and Allen found versions of the currently accepted climate model that predict a much wider range of global warming than was thought. Rather than the consensus of a 1.5° to 4.5°C increase in response to a doubling of atmospheric CO_2, some simulations run on the Oxford screen saver predict an 11°C increase, which would be catastrophic. Critics argue that such warming is unrealistic because the paleoclimate record has never revealed anything so dramatic, even in response to the

Grid Sport: Competitive Crunching

You won't find the names of Jens Seitler, Honza Cholt, John Keck, or Chris Randles among the authors of scientific papers. Nor, for that matter, the names of any of the millions of other people involved with the colossal computing projects that are predicting climate change, simulating how proteins fold, and analyzing cosmic radio data. But without their uncredited help, these projects would be nonstarters.

In the 6 years since the SETI@home screen-saver program first appeared, scientists have launched dozens of Internet projects that rely on ordinary people's computers to crunch the data while they sit idle. The result is a virtual computer that dwarfs the top supercomputer in speed and memory by orders of magnitude. The price tag? Nothing. So who are these computer philanthropists? The majority seem to be people who hear about a particular project that piques their interest, download the software, and let it run out of a sense of altruism. Others may not even be aware they are doing it. "I help about a dozen friends with repairs and upgrades to their PCs," says Christian Diepold, an English literature student from Germany, "and I install the [screen-saver software] as a kind of payment. Sometimes they don't even know it's on there."

But roughly half of the data processing contributed to these science projects comes from an entirely different sort of volunteer. They call themselves "crunchers," and they get kicks from trying to churn through more data than anyone else. As soon as the projects began publishing data-crunching statistics, competition was inevitable. Teams and rank ladders formed, and per capita crunching has skyrocketed. "I'm addicted to the stats," admits Michael, a member of a cruncher team called Rebel Alliance. To get a sense of what makes them tick, *Science* interviewed dozens of crunchers in the Internet chat forums where they socialize.

Team players. Honza Cholt says crunchers have deep discussions about the science.

Interest in crunching does not appear to correlate strongly with background. For their day jobs, hard-core crunchers are parking lot attendants, chemical engineers, stay-at-home moms and dads, insurance consultants, and even, in at least one case, miners. Their distribution, like the Internet, is global. What's the motive? People crunch "for a diversity of reasons," says Randles, a British accountant who moderates the forum for ClimatePrediction.net, but altruism tops the list. "After losing six friends over the last 2 years to cancer, I jumped at the chance to help," says an electrician in Virginia who goes by the username JTWill and runs the Find-a-Drug program on his five PCs. As a systems administrator named Josh puts it, "Why let a computer sit idle and waste electricity when you could be contributing to a greater cause?"

But another driving force is the blatant competition. Michael of Rebel Alliance has recently built a computer from scratch for the sole purpose of full-time crunching, but he says he still can't keep up with Stephen, a systems engineer in Missouri and self-proclaimed "stats junkie" who crunches on 150 computers at once. Without the competition, "it wouldn't be as much fun," says Tim, a member of Team Anandtech who crunches for Folding@home. And like any sport, rivalries are soon simmering. "Members from different teams drop in on each other's forums and taunt each other a bit," says Andy Jones, a cruncher in Northern Ireland, "but it's all in good humor." As Anandtech team member Wiz puts it, "What we have here is community."

But where does this leave the science? Do crunchers care how the fruits of their labor are used, or do they leave it all to the researchers? It depends on the project, says Cholt, a sociology student in the Czech Republic, "but the communities that form often have long and deep discussions about the science." What holds the core of the crunching community together, says Seitler, a computer specialist in Germany, is the chance "for normal people to take part in a multitude of scientific projects." In some cases, crunchers have even challenged the researchers' published conclusions. "Many scientists would groan at the thought of nonscience graduates questioning their work," says Randles, but "scrutiny beyond peer review seems an important aspect to science."

Far from indifferent, crunchers can become virtual members of the research team, says François Grey, a physicist at CERN, the particle physics lab near Geneva, Switzerland, who helps run LHC@home. Above and beyond donating their computers, "they actually help us solve problems and debug software. And you have to keep them informed about what's going on with the project, or they get upset." Crunchers might not get credited on papers, says Grey, but "scientists have to treat this community with respect."

–J.B.

CREDITS: HONZA CHOLT

Published by AAAS

largest volcanic eruptions. Stainforth emphasizes that his method does not yet allow him to attach probabilities to the different outcomes. But the upshot, he says, is that "we can't say what level of atmospheric carbon dioxide is safe." The finding runs against recent efforts to do so by politicians.

And according to Stainforth, this illustrates something that makes public-resource computing a special asset to science. Rather than a hurdle to be overcome, "public participation is half of the goal." This is particularly true for a field like climate prediction, in which the public can influence the very system being studied, but it may also be true for less political topics. "We in the SETI community have always felt that we were doing the search not just for ourselves but on behalf of all people," says Sullivan. What better way to "democratize" science than to have a research group of several million people?

–JOHN BOHANNON

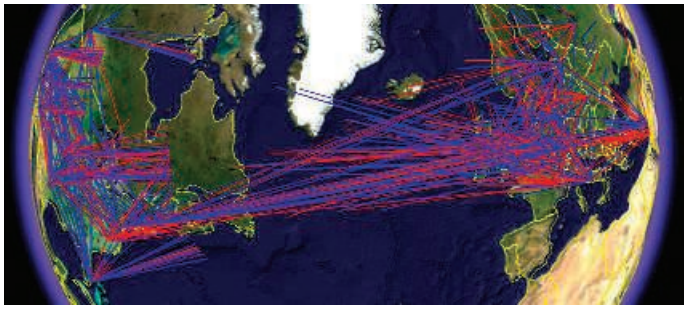
John Bohannon is a science writer based in Berlin.

NEWS

Data-Bots Chart the Internet

It's hard to map the global Internet from a small number of viewpoints. The solution may be to enlist computer users worldwide as local cartographers of cyberspace

Anyone who has tried to study the twists and turns in the data superhighway knows the problem: It is difficult even to get a decent map of the Internet. Because it grew up in a haphazard fashion with no structure imposed, no one knows how the myriad telephone lines and satellite links weave together its more than 300,000,000 computers. Today's best maps offer a badly distorted picture, incomplete and biased by a U.S. viewpoint, hampering computer scientists' efforts to design software that would make the Internet more stable and less prone to attack. But a new mapping effort may succeed where others have failed. "We want to let the Internet measure itself," says computer scientist Yuval Shavitt of Tel Aviv University in Israel, who, along with colleagues, hopes to enlist many thousands of volunteers worldwide to take part in the effort.



Gridlock. Accurate Internet maps could provide users with data traffic reports.

CREDIT: YUVAL SHAVITT AND ERAN SHIR/TEL AVIV UNIVERSITY

At the lowest level, the computers that comprise the Internet are known as "routers." They carry out the basic information housekeeping of the Net, shuttling e-mails and information packets to and fro. At a somewhat higher linked-facility level, however, the Internet can also be viewed as a network of subnetworks, or "autonomous systems," each of which corresponds to an Internet service provider or other collection of routers gathered together under a single administration. But how is this network of networks wired up?

Two years ago, computer scientist Kimberly Claffy and colleagues from the Cooperative Association for Internet Data Analysis at the University of California, San Diego, used a form of Internet "tomography" to find out. They sent out information-gathering packets from 25 computers to probe over 1 million different destinations in the Internet. Along the way, each packet recorded the links along which it moved, thereby tracing out a single path through the Internet—a chain of linked autonomous systems. Putting millions of such paths together, the researchers eventually built up a rough picture of more than 12,000 autonomous systems with more than 35,000 links between them (see www.caida.org/analysis/topology/as_core_network).

Through such efforts, researchers now understand that the Internet has a highly skewed structure, with some autonomous systems playing the role of organizing "hubs" that have far more links than most others. But researchers also know that their very best maps are still seriously incomplete.

The trouble is that all mapping efforts to date have started out from a fairly small number of sites, 50 at the most. So the maps produced tend to be biased by the locations of those sites. From some computer A, for example, researchers can send probing packets out toward computers B and C and thereby learn paths connecting A to B and A to C. But the probes would be unlikely to explore links between B and C, for the same reason that driving from New York to Boston and from New York to Montreal tells one little about the roads between Boston and Montreal. "If you send probes from only a few points, you naturally get a very partial point of view," says physicist Alessandro Vespignani, an expert on Internet topology at Indiana University, Bloomington.

To overcome this problem, Shavitt and colleagues are pioneering a new approach inspired by the idea of distributed computing. Anyone can now download a program from the Web site www.netdimes.org that will help in a global effort to map the Internet. Using no more than a few percent of the host computer's processing power, the program acts as a software agent, sending out probing packets to map local connections in and around the autonomous system in which the computer sits. "What we ask for is not so much processing power but location," says Shavitt. "We hope that the more places we have presence in, the more accurate our maps will be."

Since the project's inception late last year, individuals have downloaded nearly 800 agents that are now working together to map the Internet from 50 nations spread across all the continents. "We've already mapped out about 40,000 links between about 15,000 distinct autonomous systems, and we can already see that the Internet is about 25% denser than it was previously thought to be," says Shavitt. "This is a great project with a very new perspective," says Vespignani, who points out that better maps will help Internet administrators in predicting information bottlenecks and other hot spots.

Shavitt and his colleagues estimate that once they have about 2000 agents operating, it should be possible to get a complete map of the Internet at the autonomous-system level in less than 2 hours. Once they can do that, they hope to provide individual users with local Internet "weather reports." Ultimately, they would like to map the Internet at the level of individual routers—getting a more detailed map of the physical Internet. "We'll need about 20,000 agents distributed uniformly over the globe to get a good map at that level," says Shavitt. Then there'll be no excuse for getting lost in cyberspace.

–MARK BUCHANAN

Mark Buchanan is a writer in Cambridge, U.K.

Published by AAAS

largest volcanic eruptions. Stainforth emphasizes that his method does not yet allow him to attach probabilities to the different outcomes. But the upshot, he says, is that "we can't say what level of atmospheric carbon dioxide is safe." The finding runs against recent efforts to do so by politicians.

And according to Stainforth, this illustrates something that makes public-resource computing a special asset to science. Rather than a hurdle to be overcome, "public participation is half of the goal." This is particularly true for a field like climate prediction, in which the public can influence the very system being studied, but it may also be true for less political topics. "We in the SETI community have always felt that we were doing the search not just for ourselves but on behalf of all people," says Sullivan. What better way to "democratize" science than to have a research group of several million people?

–JOHN BOHANNON

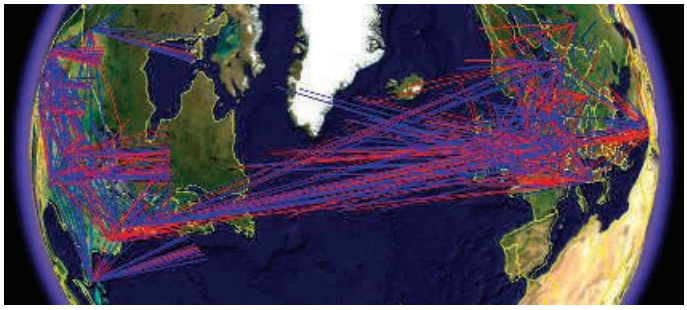
John Bohannon is a science writer based in Berlin.

NEWS

Data-Bots Chart the Internet

It's hard to map the global Internet from a small number of viewpoints. The solution may be to enlist computer users worldwide as local cartographers of cyberspace

Anyone who has tried to study the twists and turns in the data superhighway knows the problem: It is difficult even to get a decent map of the Internet. Because it grew up in a haphazard fashion with no structure imposed, no one knows how the myriad telephone lines and satellite links weave together its more than 300,000,000 computers. Today's best maps offer a badly distorted picture, incomplete and biased by a U.S. viewpoint, hampering computer scientists' efforts to design software that would make the Internet more stable and less prone to attack. But a new mapping effort may succeed where others have failed. "We want to let the Internet measure itself," says computer scientist Yuval Shavitt of Tel Aviv University in Israel, who, along with colleagues, hopes to enlist many thousands of volunteers worldwide to take part in the effort.



Gridlock. Accurate Internet maps could provide users with data traffic reports.

CREDIT: YUVAL SHAVITT AND ERAN SHIR/TEL AVIV UNIVERSITY

At the lowest level, the computers that comprise the Internet are known as "routers." They carry out the basic information housekeeping of the Net, shuttling e-mails and information packets to and fro. At a somewhat higher linked-facility level, however, the Internet can also be viewed as a network of subnetworks, or "autonomous systems," each of which corresponds to an Internet service provider or other collection of routers gathered together under a single administration. But how is this network of networks wired up?

Two years ago, computer scientist Kimberly Claffy and colleagues from the Cooperative Association for Internet Data Analysis at the University of California, San Diego, used a form of Internet "tomography" to find out. They sent out information-gathering packets from 25 computers to probe over 1 million different destinations in the Internet. Along the way, each packet recorded the links along which it moved, thereby tracing out a single path through the Internet—a chain of linked autonomous systems. Putting millions of such paths together, the researchers eventually built up a rough picture of more than 12,000 autonomous systems with more than 35,000 links between them (see www.caida.org/analysis/topology/as_core_network).

Through such efforts, researchers now understand that the Internet has a highly skewed structure, with some autonomous systems playing the role of organizing "hubs" that have far more links than most others. But researchers also know that their very best maps are still seriously incomplete.

The trouble is that all mapping efforts to date have started out from a fairly small number of sites, 50 at the most. So the maps produced tend to be biased by the locations of those sites. From some computer A, for example, researchers can send probing packets out toward computers B and C and thereby learn paths connecting A to B and A to C. But the probes would be unlikely to explore links between B and C, for the same reason that driving from New York to Boston and from New York to Montreal tells one little about the roads between Boston and Montreal. "If you send probes from only a few points, you naturally get a very partial point of view," says physicist Alessandro Vespignani, an expert on Internet topology at Indiana University, Bloomington.

To overcome this problem, Shavitt and colleagues are pioneering a new approach inspired by the idea of distributed computing. Anyone can now download a program from the Web site www.netdimes.org that will help in a global effort to map the Internet. Using no more than a few percent of the host computer's processing power, the program acts as a software agent, sending out probing packets to map local connections in and around the autonomous system in which the computer sits. "What we ask for is not so much processing power but location," says Shavitt. "We hope that the more places we have presence in, the more accurate our maps will be."

Since the project's inception late last year, individuals have downloaded nearly 800 agents that are now working together to map the Internet from 50 nations spread across all the continents. "We've already mapped out about 40,000 links between about 15,000 distinct autonomous systems, and we can already see that the Internet is about 25% denser than it was previously thought to be," says Shavitt. "This is a great project with a very new perspective," says Vespignani, who points out that better maps will help Internet administrators in predicting information bottlenecks and other hot spots.

Shavitt and his colleagues estimate that once they have about 2000 agents operating, it should be possible to get a complete map of the Internet at the autonomous-system level in less than 2 hours. Once they can do that, they hope to provide individual users with local Internet "weather reports." Ultimately, they would like to map the Internet at the level of individual routers—getting a more detailed map of the physical Internet. "We'll need about 20,000 agents distributed uniformly over the globe to get a good map at that level," says Shavitt. Then there'll be no excuse for getting lost in cyberspace.

–MARK BUCHANAN

Mark Buchanan is a writer in Cambridge, U.K.

Published by AAAS

Service-Oriented Science

Ian Foster

New information architectures enable new approaches to publishing and accessing valuable data and programs. So-called service-oriented architectures define standard interfaces and protocols that allow developers to encapsulate information tools as services that clients can access without knowledge of, or control over, their internal workings. Thus, tools formerly accessible only to the specialist can be made available to all; previously manual data-processing and analysis tasks can be automated by having services access services. Such service-oriented approaches to science are already being applied successfully, in some cases at substantial scales, but much more effort is required before these approaches are applied routinely across many disciplines. Grid technologies can accelerate the development and adoption of service-oriented science by enabling a separation of concerns between discipline-specific content and domain-independent software and hardware infrastructure.

Paul Erdös claimed that a mathematician is a machine for turning coffee into theorems. The scientist is arguably a machine for turning data into insight. However, advances in information technology are changing the way in which this role is fulfilled—by automating time-consuming activities and thus freeing the scientist to perform other tasks. In this Viewpoint, I discuss how service-oriented computing—technology that allows powerful information tools to be made available over the network, always on tap, and easy for scientists to use—may contribute to that evolution.

The practice of science has, of course, already been affected dramatically by information technology and, in particular, by the Internet. For example, the hundreds of gigabytes of genome sequence available online means that for a growing number of biologists, "data" is something that they find on the Web, not in the lab. Similarly, emerging "digital observatories" [already several hundred terabytes in dozens of archives (*1*)] allow astronomers to pose and answer in seconds questions that might previously have required years of observation. In fields such as cosmology and climate, supercomputer simulations have emerged as essential tools, themselves producing large data sets that, when published online, are of interest to many (*2*). An exploding number of sensors (*3*), the rapidly expanding computing and storage capabilities of federated Grids (*4*), and advances in optical networks (*5*) are accelerating these trends by making increasingly powerful capabilities available online.

Sometimes, however, the thrill of the Web seems to blind us to the true implications of these developments. Human access to online resources is certainly highly useful, putting a global library at our fingertips. But ultimately, it is automated access by software programs that will be truly revolutionary, simply because of the higher speeds at which programs can operate. In the time that a human user takes to locate one useful piece of information within a Web site, a program may access and integrate data from many sources and identify relationships that a human might never discover unaided. Two dramatic examples are systems that automatically integrate information from genome and protein sequence databases to infer metabolic pathways (*6*) and systems that search digital sky surveys to locate brown dwarfs (*7*).

The key to such success is uniformity of interface, so that programs can discover and access services without the need to write custom code for each specific data source, program, or sensor. Electric power–transmission standards and infrastructure enabled development of the electric power grid and spurred the development of a plethora of electric tools. In a similar manner, service technologies enable the development of a wide range of programs that integrate across multiple existing services for purposes such as metabolic pathway reconstruction, categorization of astronomical objects, and analysis of environmental data. If such programs are themselves made accessible as services, the result can be the creation of distributed networks of services, each constructed by a different individual or group, and each providing some original content and/or value-added product (*8*).

We see this evolution occurring in the commercial Internet. As the Web has expanded in scale, so the preferred means of finding things has evolved from Yahoo's manually assembled lists to Google's automatically computed indices. Now Google is making its indices accessible, spurring development of yet other services. What makes Google's indices feasible is the existence of large quantities of data in a uniform format (HTML, HyperText Markup Language) and—two important factors that must be considered when we turn to science—smart computer scientists to develop the algorithms and software required to manage the 100,000 computers used (at last count) to analyze Web link structure, and smart businesspeople to raise the money that pays for those computers!

The term "service-oriented architecture" refers to systems structured as networks of loosely coupled, communicating services (*9*). Thus, "service-oriented science" refers to scientific research enabled by distributed networks of interoperating services. [The term "e-Science," coined by John Taylor, has a similar but broader connotation (*10*).]

Creating and Sharing Services

Creating a service involves describing, in some conventional manner, the operations that the service supports; defining the protocol used to invoke those operations over the Internet; and operating a server to process incoming requests. A set of technologies called Web services (*9*) are gaining wide acceptance for these purposes. A variety of commercial and open-source Web services tools exist for developing services, deploying and operating services, and developing client applications. A fair amount of experience has been gained with the creation of services and applications in different science domains. Although problems remain (e.g., efficiency, interoperability of different vendor offerings), the technology is well beyond the experimental stage. Nevertheless, it can still be a big step to realize the full potential of service-oriented science, for reasons that I now discuss.

Interoperability. Services have little value if others cannot discover, access, and make sense of them. Yet, as Stein has observed (*11*), today's scientific communities too often resemble medieval Italy's collection of warring city states, each with its own legal system and dialect. Web services mechanisms for describing, discovering, accessing, and securing services provide a common alphabet, but a true lingua franca requires agreement on protocols, data formats, and ultimately semantics (*12*). For example, the definition of VOTable, a standard XML (eXtensible Markup Language)–based representation for tabular data (*13*), has been a powerful force for progress in astronomy.

Scale. Services must often deal with data volumes, computational demands, and numbers of users beyond the capacity of a typical PC. Responding to a user request—or to the arrival of new data—can involve large amounts of computation. For example, the Argonne GNARE system searches periodically through DNA and protein databases for new and updated genomes and then computes and pub-

Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA, and Department of Computer Science, University of Chicago, Chicago, IL 60637, USA. E-mail: foster@mcs.anl.gov

lishes derived values (*14*) (Fig. 1). Analysis of a single bacterial genome of 4000 sequences by three bioinformatics tools (BLAST, PFAM, and BLOCKS) requires 12,000 steps, each taking on the order of 30 s of run time. GNARE is able to perform these tasks in a timely fashion only because it has access to distributed resources provided by two U.S. national-scale infrastructures, TeraGrid and Open Science Grid (see below).

The impact of automation on service load must also be considered. It is improbable that even a tiny fraction of the perhaps 500,000 biologists worldwide will decide to access GenBank, GNARE, or any other service at the same time. However, it is quite conceivable that 50,000 "agents" operating on their behalf would do so—and that each such agent would want to generate thousands of requests.

Management. In a networked world, any useful service will become overloaded. Thus, we need to control who uses services and for what purposes. Particularly valuable services may become community resources requiring coordinated management. Grid architectures and software—a set of Web services technologies focused on distributed system management—can play an important role in this regard (*15*).

Quality control. As the number and variety of services grow and interdependencies among services increase, it becomes important to automate previously manual quality-control processes—so that, for example, users can determine the provenance of a particular derived data product (*8*, *16*). The ability to associate metadata with data and services can be important, as can the ability to determine the identity of entities that assert metadata, so that consumers can make their own decisions concerning quality.

Incentives. A scientist may work long hours in the lab to obtain results that may bring tenure, fame, or fortune. The same time spent developing a service may not be so rewarded. We need to change incentives and enable specialization so that being a service developer is as honorable as being an experimentalist or theorist. Intellectual property issues must also be addressed so that people feel comfortable making data available freely. It is perhaps not surprising that astronomy has led the way in putting data online, given that its data have no known commercial value.

Scientists are certainly not alone in grappling with these challenges. However, science is perhaps unique in the scope and scale of its problems, the number and diversity of potential contributors, and the subtlety of the questions that service networks can be used to answer.

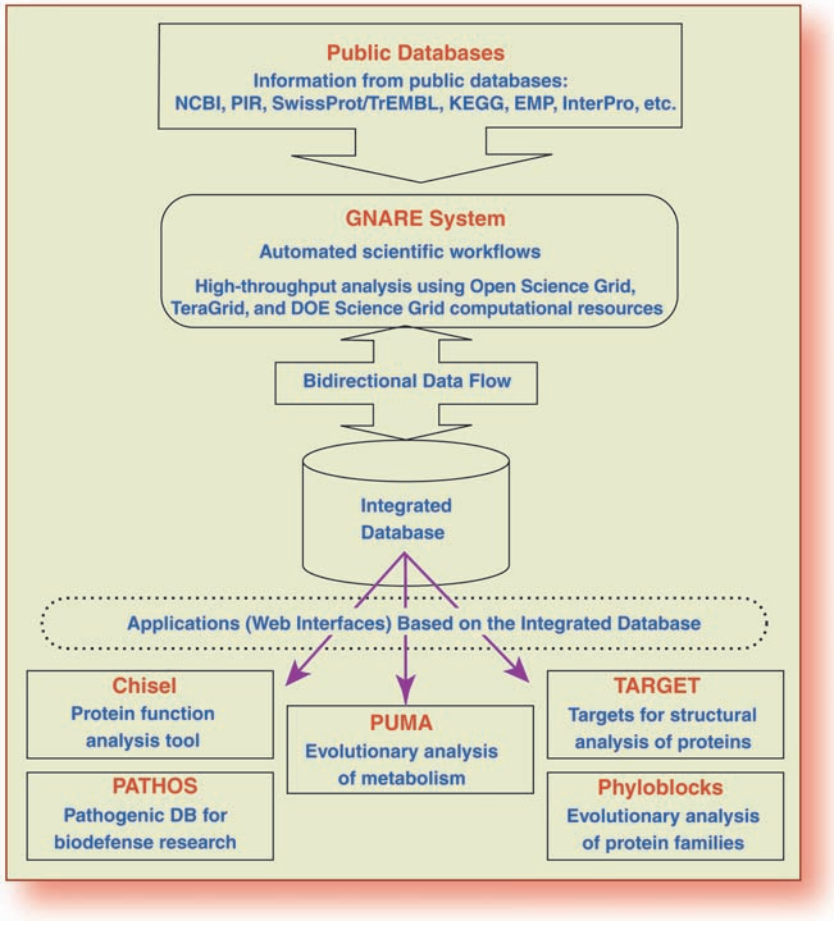


Fig. 1. What it can take to build a service. A powerful approach to the interpretation of newly sequenced genomes is comparative analysis against all annotated sequences in publicly available resources. Currently, the largest sequence database at the National Center for Biotechnology Information contains 2.3 million protein sequences. The precision of genetic sequence analysis and assignment of function to genes can be increased markedly by the use of multiple bioinformatics algorithms for data analysis. The GNARE system discussed in the text precomputes analysis results for every sequence, finding protein similarities (BLAST), protein family domains (BLOCKS), and structural characteristics. Grid resources are used to run the resulting millions of processes, a task that must be repeated frequently owing to the exponentially growing amount of data. [Image credit: Bioinformatics group, Mathematics and Computer Science Division, Argonne National Laboratory]

Rethinking Infrastructure

As scale increases, creating, operating, and even accessing services become increasingly challenging. How do we ensure that service-oriented science realizes its promise of being a democratizing force, rather than increasing the gap between the "haves" and "have-nots"?

Part of the solution is a familiar idea in commercial information technology, namely, outsourcing. Building and deploying a service require expertise and resources in three distinct areas: (i) the domain-specific content—data, software, and processes—that is to be shared; (ii) the domain-independent software functions needed to operate and manage the service and to enable community access, such as membership services, registries, metadata catalogs, and workflow orchestration services; and (iii) the physical resources—networks, storage, and computers—needed to host content and functions.

The last two capabilities—functions and resources—can, in principle, be handed off to specialist providers. If such specialists can deliver resources or operate required functions for many communities, then (again, in principle) economies of scale can be achieved, while scientists can focus on what they are good at—providing content and advancing science. In addition, individual services can scale more easily and efficiently when needed.

To see how this strategy can work, consider the SourceForge system, which provides hosting capabilities for communities developing open-source software. A new open-source project is provided with access to code archiving, mailing lists, and other related functions, as well as the hardware required to host those functions. This outsourcing of function and resource is made possible by the existence of the Internet infrastructure along with standard Web servers, browsers, and associated protocols, which together allow users (in this case, open-source communities) to focus on providing content (code) while SourceForge runs Web servers and related infrastructure.

In a similar manner, a "SourceForge for science" would both host scientific communities—operating community membership services, catalogs, storage services, work-

Fig. 2. The Open Science Grid links storage and computing resources at more than 30 sites across the United States to support a variety of services and applications, many concerned with large-scale data analysis. Circles show a subset of Open Science Grid sites; lines indicate communications, some with international partners. [Image credit: I. Legrand, Caltech]

flow orchestration services, and so forth—and provide access to the hardware resources required to operate both those functions and the application-specific services that constitute the communities "content." In this case, the supporting infrastructure must provide a much richer set of capabilities than does SourceForge, encompassing, for example, access control, accounting, provisioning, and related management issues. As noted above, Grid architectures and software (*15*) address many of these concerns, allowing users to focus on providing "content," which in this case comprises not just Web pages but also services, data, and programs.

SourceForge's hardware requirements are not substantial and thus can easily be provided by a centralized system. However, "cyberinfrastructure" (*17*) to support scientific communities need not be centralized. For example, the Open Science Grid (OSG) collaboration has constructed a distributed "Grid" linking clusters at 30 sites across the United States that total thousands of computers and tens of terabytes of storage (*18*) (Fig. 2). The Enabling Grids for eScience in Europe project, EGEE, has a similar structure. Major research universities and national laboratories participate in OSG and EGEE, but so do smaller institutions, which can thus enhance educational and research opportunities. For example, Florida International University is an important OSG resource provider, thanks to its 92-processor Linux cluster. All participants can obtain access to large quantities of distributed storage and computational power when they need it. These systems are being used by researchers in high-energy physics, biology, chemistry, radiology, and computer science.

This separation of concerns also suggests new roles for campus information technology organizations. In addition to operating commodity services such as Internet and e-mail, these organizations can host functions and provide resources.

Approaches to Scaling

The many groups working to apply service-oriented techniques to science are each exploring one or more of three different approaches to the problem of scaling. In the first, "cookie-cutter" approach, researchers create dedicated domain-specific infrastructures, in which uniformity is enforced across the board, at the content, function, and resource level. Here, the community standardizes the domain-specific software—and often also the hardware—that participants must deploy in order to provide required functions and resources. I give three examples of such systems.

The Biomedical Informatics Research Network, BIRN (*19*), is a National Institutes of Health initiative to facilitate collaboration in the biomedical sciences. BIRN has deployed standard compute and storage clusters at 19 sites across the United States. These systems, plus various functions such as catalogs and ontologies, support a variety of collaborative research programs in areas such as mouse brain morphology (*20*).

The National Science Foundation's Network for Earthquake Engineering Simulation, NEES, is a national collaboratory enabling community access to specialized instrument, data, and simulation resources for earthquake engineering. Each of its 17 instrument sites runs a NEES Point of Presence (a modest PC with a standard hardware configuration) with standard software enabling teleobservation, teleoperation, data collection, and related functions. Central services include catalogs and data archives. NEES has already enabled unique distributed experiments involving facilities at multiple sites (*21*).

The PlanetLab computer science testbed is a collection of several hundred PCs at universities and research laboratories worldwide, each with a standard configuration and each running standard software (*22*). Computer scientists can obtain access to "slices" on distributed collections of these PCs, on which they can deploy and evaluate experimental distributed services.

Pushing the electric power grid analogy perhaps farther than we should, cookie-cutter approaches give each participant their own electricity generator. This strategy has the advantage of achieving a high degree of central control and thus uniformity. On the other hand, the cost of expanding capability is high, requiring the acquisition and deployment of new hardware.

In the second, more bottom-up approach, researchers develop service ecologies in which agreements on interfaces allow participants to provide content and function in any way they see fit.

I referred above to the international virtual observatory community's VOTable format and to work in bioinformatics. The Department of Energy's Earth System Grid, ESG (*2*), is another example of a discipline-specific service that emphasizes the definition and implementation of standard interfaces. Building on the widely used OPeNDAP protocol for publishing and accessing environmental data, ESG has deployed services that provide access to over 100 TB of climate simulation data from the National Center for Atmospheric Research's Community Climate Simulation Model and other models involved in the International Panel on Climate Change assessment. Many terabytes of data are downloaded from these services each month.

As a second example, the UK myGrid project (*8*) has developed tools that allow biologists to define workflows that integrate information from multiple sources, including both biological databases and bioinformatics applications. These workflows can be archived and then run periodically to identify new phenomena of interest as, for example, in a recent study of Williams-Beuren syndrome (*23*).

For a third example, the Department of Energy's Fusion Collaboratory (*24*) operates services that enable online access to simulation codes. By reducing barriers to use, these services are increasing use of advanced computational techniques. Project members have also demonstrated on-demand coupling of simulation capabilities with physical experiments.

Continuing the electric power grid analogy, such service ecologies define relevant standards but leave each site to acquire and configure its own equipment. This approach has the advantage that the cost of entry can be low, particularly if appropriate software is available. On the other hand, individual service providers have no immediate means of scaling capability beyond acquiring more hardware.

The third approach involves the definition and deployment of general-purpose infrastructures that deliver discipline-independent resources or functions. I have already mentioned OSG and EGEE. As a third example, the National Science Foundation's TeraGrid links resources at nine sites across the United States, with each site deploying a common software distribution that permits secure remote access to computers and storage systems, monitoring of system components, accounting for usage, and so on. TeraGrid targets not only high-end "power users" but also the larger community through the deployment of "science gateways," discipline-specific services hosted on TeraGrid in support of specific communities.

General-purpose infrastructures can be compared with power plants, which operate to provide electricity to any consumer connected to the electric power grid. Like power plants, they have the potential to achieve economies of scale but also must grapple with the challenges of supporting many users with diverse requirements.

In addition to these national or transnational efforts, many university campuses are deploying "campus Grids" to support faculty and students. For example, Purdue University's NanoHub provides students and faculty with access to various applications, while the UCLA Grid federates multiple clusters across campus and provides online access to popular simulation codes.

These projects, and many others like them, are important experiments in the policies, organizational structures, and mechanisms required to realize service-oriented science. Elements of all three approaches will be required if we are to achieve broad adoption. In particular, it cannot be efficient for every scientist and community to become a service provider. Instead, individual communities—especially smaller communities—should be able to outsource selected functions and physical resources, thus allowing them to focus on developing their domain-specific content. The successful creation and operation of the service providers that support this outsourcing require both Grid infrastructure software and organizational and funding structures that expose real costs so that "build versus buy" decisions can be made in an informed manner.

Summary

Service-oriented science has the potential to increase individual and collective scientific productivity by making powerful information tools available to all, and thus enabling the widespread automation of data analysis and computation. Ultimately, we can imagine a future in which a community's shared understanding is no longer documented exclusively in the scientific literature but is documented also in the various databases and programs that represent—and automatically maintain and evolve—a collective knowledge base.

Service-oriented science is also a new way of doing business, with implications for all aspects of the scientific enterprise. Students and researchers must acquire new skills to build and use services. New cyberinfrastructure is required to host services, especially as demand increases. Policies governing access to services must evolve. Above all, much hard work must be done in both disciplinary science and information technology in order to develop the understanding needed for this potential to be fully exploited.

References and Notes

1. A. Szalay, J. Gray, *Science* **293**, 2037 (2001).
2. D. Bernholdt *et al.*, *Proc. IEEE* **93**, 485 (2005).
3. D. Culler, H. Mulder, *Sci. Am.* **290**, 84 (June 2004).
4. I. Foster, *Sci. Am.* **288**, 78 (April 2003).
5. T. DeFanti, C. de Laat, J. Mambretti, K. Neggers, B. St Arnaud, *Commun. ACM* **46**, 34 (2003).
6. R. Overbeek *et al.*, *Nucleic Acids Res.* **28**, 123 (2000).
7. Z. Tsvetanov *et al.*, *Astrophys. J.* **531**, L61 (2000).
8. C. Goble, S. Pettifer, R. Stevens, in *The Grid: Blueprint for a New Computing Infrastructure* (Morgan Kaufmann, San Francisco, ed. 2, 2004), pp. 121–134.
9. D. Booth *et al.*, *Web Services Architecture* (W3C, Working Draft, 2003; www.w3.org/TR/2003/WD-ws-arch-20030808).
10. T. Hey, A. E. Trefethen, *Science* **308**, 817 (2005).
11. L. Stein, *Nature* **317**, 119 (2002).
12. T. Berners-Lee, J. Hendler, O. Lassila, *Sci. Am.* **284**, 34 (May 2001).
13. F. Ochsenbein *et al.*, *VOTable Format Definition Version 1.1* (International Virtual Observatory Alliance, 2004; www.ivoa.net/Documents/latest/VOT.html).
14. D. Sulakhe *et al.*, *J. Clin. Monitor. Comput.*, in press.
15. I. Foster, C. Kesselman, J. Nick, S. Tuecke, *IEEE Comput.* **35**, 37 (2002).
16. J. Myers, A. Chappell, M. Elder, A. Geist, J. Schwidder, *IEEE Comput. Sci. Eng.* **5**, 44 (2003).
17. "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure" (National Science Foundation, Washington, DC, 2003; www.communitytechnology.org/nsf_ci_report).
18. I. Foster *et al.*, in *13th IEEE International Symposium on High Performance Distributed Computing* (IEEE, Los Alamitos, CA, 2004), pp. 236–245.
19. M. Ellisman, S. Peltier, in *The Grid: Blueprint for a New Computing Infrastructure* (Morgan Kaufmann, San Francisco, ed. 2, 2004), pp. 109–120.
20. M. Martone, A. Gupta, M. Ellisman, *Nat. Neurosci.* **7**, 467 (2004).
21. L. Pearlman *et al.*, in *13th IEEE Intl. Symp. on High Performance Distributed Computing* (IEEE, Los Alamitos, CA, 2004), pp. 14–23.
22. A. Bavier *et al.*, in *1st Symposium on Networked Systems Design and Implementation* (Usenix, Berkeley, CA, 2004), pp. 253–266.
23. R. Stevens *et al.*, *Bioinformatics* **20**, (suppl. 1), i303 (2004).
24. K. Keahey *et al.*, *Future Generation Comput. Syst.* **18**, 1005 (2002).

25. I gratefully acknowledge helpful discussions with C. Catlett, C. Kesselman, M. Livny, A. Szalay, and R. Stevens. This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-Eng-38, and by the NSF. The author is Founder of and Chief Open Source Strategist at Univa Corporation.

10.1126/science.1110411

VIEWPOINT

Cyberinfrastructure for e-Science

Tony Hey and Anne E. Trefethen

Here we describe the requirements of an e-Infrastructure to enable faster, better, and different scientific research capabilities. We use two application exemplars taken from the United Kingdom's e-Science Programme to illustrate these requirements and make the case for a service-oriented infrastructure. We provide a brief overview of the UK "plug-and-play composable services" vision and the role of semantics in such an e-Infrastructure.

It is no coincidence that it was at CERN, the particle physics accelerator laboratory in Geneva, that Tim Berners-Lee invented the World Wide Web. Given the distributed nature of the multi-institute collaborations required for modern particle physics experiments, researchers desperately needed a tool for exchanging information. After a slow start, the community enthusiastically adopted the Web for information exchange within their global experimental collaborations. Since its beginnings in the early 1990s, the Web has become an indispensable tool not just for the scientific world, but for the humanities, business, and recreation. Now, just a decade later, scientists are attempting to develop capabilities for collaboration that go far beyond those of the Web. Besides being able to access information from different sites, they want to be able to integrate, federate, and analyze infor-

Engineering and Physical Sciences Research Council, Polaris House, North Star Avenue, Swindon SN2 1ET, UK.

Continuing the electric power grid analogy, such service ecologies define relevant standards but leave each site to acquire and configure its own equipment. This approach has the advantage that the cost of entry can be low, particularly if appropriate software is available. On the other hand, individual service providers have no immediate means of scaling capability beyond acquiring more hardware.

The third approach involves the definition and deployment of general-purpose infrastructures that deliver discipline-independent resources or functions. I have already mentioned OSG and EGEE. As a third example, the National Science Foundation's TeraGrid links resources at nine sites across the United States, with each site deploying a common software distribution that permits secure remote access to computers and storage systems, monitoring of system components, accounting for usage, and so on. TeraGrid targets not only high-end "power users" but also the larger community through the deployment of "science gateways," discipline-specific services hosted on TeraGrid in support of specific communities.

General-purpose infrastructures can be compared with power plants, which operate to provide electricity to any consumer connected to the electric power grid. Like power plants, they have the potential to achieve economies of scale but also must grapple with the challenges of supporting many users with diverse requirements.

In addition to these national or transnational efforts, many university campuses are deploying "campus Grids" to support faculty and students. For example, Purdue University's NanoHub provides students and faculty with access to various applications, while the UCLA Grid federates multiple clusters across campus and provides online access to popular simulation codes.

These projects, and many others like them, are important experiments in the policies, organizational structures, and mechanisms required to realize service-oriented science. Elements of all three approaches will be required if we are to achieve broad adoption. In particular, it cannot be efficient for every scientist and community to become a service provider. Instead, individual communities—especially smaller communities—should be able to outsource selected functions and physical resources, thus allowing them to focus on developing their domain-specific content. The successful creation and operation of the service providers that support this outsourcing require both Grid infrastructure software and organizational and funding structures that expose real costs so that "build versus buy" decisions can be made in an informed manner.

Summary

Service-oriented science has the potential to increase individual and collective scientific productivity by making powerful information tools available to all, and thus enabling the widespread automation of data analysis and computation. Ultimately, we can imagine a future in which a community's shared understanding is no longer documented exclusively in the scientific literature but is documented also in the various databases and programs that represent—and automatically maintain and evolve—a collective knowledge base.

Service-oriented science is also a new way of doing business, with implications for all aspects of the scientific enterprise. Students and researchers must acquire new skills to build and use services. New cyberinfrastructure is required to host services, especially as demand increases. Policies governing access to services must evolve. Above all, much hard work must be done in both disciplinary science and information technology in order to develop the understanding needed for this potential to be fully exploited.

References and Notes

1. A. Szalay, J. Gray, *Science* **293**, 2037 (2001).
2. D. Bernholdt *et al.*, *Proc. IEEE* **93**, 485 (2005).
3. D. Culler, H. Mulder, *Sci. Am.* **290**, 84 (June 2004).
4. I. Foster, *Sci. Am.* **288**, 78 (April 2003).
5. T. DeFanti, C. de Laat, J. Mambretti, K. Neggers, B. St Arnaud, *Commun. ACM* **46**, 34 (2003).
6. R. Overbeek *et al.*, *Nucleic Acids Res.* **28**, 123 (2000).
7. Z. Tsvetanov *et al.*, *Astrophys. J.* **531**, L61 (2000).
8. C. Goble, S. Pettifer, R. Stevens, in *The Grid: Blueprint for a New Computing Infrastructure* (Morgan Kaufmann, San Francisco, ed. 2, 2004), pp. 121–134.
9. D. Booth *et al.*, *Web Services Architecture* (W3C, Working Draft, 2003; www.w3.org/TR/2003/WD-ws-arch-20030808).
10. T. Hey, A. E. Trefethen, *Science* **308**, 817 (2005).
11. L. Stein, *Nature* **317**, 119 (2002).
12. T. Berners-Lee, J. Hendler, O. Lassila, *Sci. Am.* **284**, 34 (May 2001).
13. F. Ochsenbein *et al.*, *VOTable Format Definition Version 1.1* (International Virtual Observatory Alliance, 2004; www.ivoa.net/Documents/latest/VOT.html).
14. D. Sulakhe *et al.*, *J. Clin. Monitor. Comput.*, in press.
15. I. Foster, C. Kesselman, J. Nick, S. Tuecke, *IEEE Comput.* **35**, 37 (2002).
16. J. Myers, A. Chappell, M. Elder, A. Geist, J. Schwidder, *IEEE Comput. Sci. Eng.* **5**, 44 (2003).
17. "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue Ribbon Advisory Panel on Cyberinfrastructure" (National Science Foundation, Washington, DC, 2003; www.communitytechnology.org/nsf_ci_report).
18. I. Foster *et al.*, in *13th IEEE International Symposium on High Performance Distributed Computing* (IEEE, Los Alamitos, CA, 2004), pp. 236–245.
19. M. Ellisman, S. Peltier, in *The Grid: Blueprint for a New Computing Infrastructure* (Morgan Kaufmann, San Francisco, ed. 2, 2004), pp. 109–120.
20. M. Martone, A. Gupta, M. Ellisman, *Nat. Neurosci.* **7**, 467 (2004).
21. L. Pearlman *et al.*, in *13th IEEE Intl. Symp. on High Performance Distributed Computing* (IEEE, Los Alamitos, CA, 2004), pp. 14–23.
22. A. Bavier *et al.*, in *1st Symposium on Networked Systems Design and Implementation* (Usenix, Berkeley, CA, 2004), pp. 253–266.
23. R. Stevens *et al.*, *Bioinformatics* **20**, (suppl. 1), i303 (2004).
24. K. Keahey *et al.*, *Future Generation Comput. Syst.* **18**, 1005 (2002).

25. I gratefully acknowledge helpful discussions with C. Catlett, C. Kesselman, M. Livny, A. Szalay, and R. Stevens. This work was supported in part by the Mathematical, Information, and Computational Sciences Division subprogram of the Office of Advanced Scientific Computing Research, Office of Science, U.S. Department of Energy, under Contract W-31-109-Eng-38, and by the NSF. The author is Founder of and Chief Open Source Strategist at Univa Corporation.

10.1126/science.1110411

VIEWPOINT

Cyberinfrastructure for e-Science

Tony Hey and Anne E. Trefethen

Here we describe the requirements of an e-Infrastructure to enable faster, better, and different scientific research capabilities. We use two application exemplars taken from the United Kingdom's e-Science Programme to illustrate these requirements and make the case for a service-oriented infrastructure. We provide a brief overview of the UK "plug-and-play composable services" vision and the role of semantics in such an e-Infrastructure.

It is no coincidence that it was at CERN, the particle physics accelerator laboratory in Geneva, that Tim Berners-Lee invented the World Wide Web. Given the distributed nature of the multi-institute collaborations required for modern particle physics experiments, researchers desperately needed a tool for exchanging information. After a slow start, the community enthusiastically adopted the Web for information exchange within their global experimental collaborations. Since its beginnings in the early 1990s, the Web has become an indispensable tool not just for the scientific world, but for the humanities, business, and recreation. Now, just a decade later, scientists are attempting to develop capabilities for collaboration that go far beyond those of the Web. Besides being able to access information from different sites, they want to be able to integrate, federate, and analyze infor-

Engineering and Physical Sciences Research Council, Polaris House, North Star Avenue, Swindon SN2 1ET, UK.

mation from many disparate and distributed data sources (including data archives as well as networks of sensors and identification tags) and to access and control computing resources and experimental equipment at remote sites. Such an infrastructure is in fact very close to the vision of linking computers and accessing remote data that J. C. R. Licklider took with him to the Defense Advanced Research Projects Agency, which initiated the research project that led to the ARPANET (which later became the Internet) (*1*).

One of the key drivers behind the search for such new scientific tools is the imminent deluge of data from new generations of scientific experiments and surveys (*2*). In order to exploit and explore the petabytes of scientific data that will arise from these high-throughput experiments, supercomputer simulations, sensor networks, and satellite surveys, scientists will need assistance from specialized search engines, data mining tools, and data visualization tools that make it easy to ask questions and understand answers. To create such tools, the data will need to be annotated with relevant "metadata" giving information as to provenance, content, conditions, and so on; and, in many instances, the sheer volume of data will dictate that this process be automated. Scientists will create vast distributed digital repositories of scientific data requiring management services similar to those of more conventional digital libraries, as well as other data-specific services. The ability to search, access, move, manipulate, and mine such data will be a central requirement for this new generation of collaborative science software applications.

In the United Kingdom, this vision was articulated by John Taylor, then director general of Research Councils at the Office of Science and Technology (OST)—a position roughly equivalent to that of the director of the National Science Foundation (NSF) in the United States. Taylor came from Hewlett-Packard, which has long had a vision of utility computing in which users in the future would be able to pay for information technology (IT) services as they required them, in the same way as we use conventional utilities such as electricity, gas, and water, or in pay-as-you-go telephone billing, rather than purchase IT infrastructure outright. Taylor recognized the trends in scientific collaboration summarized above and realized that many areas of science could benefit from a common IT infrastructure to support multidisciplinary and distributed collaborations. He therefore put together a successful bid to the UK government (*3*), and in 2001 the United Kingdom initiated a £250 million, 5-year e-Science program to develop the tools, technologies, and infrastructure to support such multidisciplinary and collaborative science. It is important to emphasize that e-Science is not a new scientific discipline; rather, the e-Science infrastructure developed by the program should allow scientists to do faster, better, or different research. This claim is best illustrated by two examples.

Two e-Science Exemplars

The global particle physics community is now planning a series of experiments to find the hitherto elusive Higgs boson. This particle is a key component of the successful Standard Model of Glashow, Salam, and Weinberg that is believed to unify the weak and electromagnetic interactions (*4*). At the CERN laboratory in Geneva, the world's most powerful particle accelerator—the Large Hadron Collider (LHC)—is under construction and is scheduled to be operational by 2007. However, finding experimental evidence for the existence of the Higgs particle will be a major technological challenge, because the characteristic signals of the Higgs are expected to be very rare and subtle. Experiments at the LHC will be on a scale greater than any other previous physics experiments, and each will generate several petabytes of data per year. The major experiments are collaborations of over 1000 physicists from over 100 institutions in Europe, America, and Asia. The experimental data, although initially generated at CERN, are distributed to groups of scientists all over the world. Not all of the analysis can be done in Geneva. Thus, very large amounts of data will need to be routinely distributed for subsequent analysis by teams of physicists at the collaborating institutions. In addition to the large volumes of experimental data, the particle physicists in each experiment will also create large samples of simulated data in order to understand the detailed behavior of the experimental detectors. The e-Science infrastructure required for these LHC experiments goes far beyond the capability to access data on static Web sites. The experimental particle physicists are therefore building a global infrastructure—the LHC Computing Grid—that will permit the transport and data mining of huge distributed data sets (*5*). This "middleware" infrastructure (so called because the software lies between the network and the application) will enable physicists to set up appropriate data sharing/replication/management services and to facilitate decentralized computational simulations and analysis.

A second and perhaps more typical example of multidisciplinary collaborative science is in the emerging field of systems biology. The UK program has recently funded a major e-Science project on Integrative Biology (*6*). This is a £2.3 million project led by Oxford University, whose goal is to develop a virtual laboratory for research on heart disease and cancer. The project involves four other UK universities, together with the University of Auckland in New Zealand. Denis Noble's group at Oxford are world-renowned for their research into models of the electrical behavior of heart cells. Peter Hunter and his team in the bioengineering department at the University of Auckland in New Zealand are doing pioneering research into mechanical models of the beating heart. Both groups are currently doing world-class research in their own specialist areas. However, the project aims to connect researchers in these two groups in a scientific virtual organization (VO). This VO is an environment that will allow researchers in the project (and only researchers in the project) to routinely access the models and data developed at both Oxford and Auckland, as well as allowing them access to computing resources and UK supercomputers. Of course, researchers have long been able to access resources at a remote site; here the intent is to put in place a comprehensive infrastructure that can provide users with a single sign-on capability that authenticates each user and authorizes access to specific resources at each site, automatically negotiating problems with firewalls and multiple administrative authorities. By providing a powerful and usable e-Science research environment in which these two groups can combine their research activities, it will be possible to investigate links between specific gene defects that affect the electrical behavior of heart cells and life-threatening heart arrhythmias. This is a type of research that neither group can do independently; it is in this sense that e-Science technologies can enable different science.

Cyberinfrastructure, e-Infrastructure, and the Grid

The high-speed national research networks that constitute the underlying fabric of the academic Internet have long connected scientific collaborations such as these. But now under the banner of e-Science, scientists and computer scientists around the world are collaborating to construct a set of software tools and services to be deployed on top of these physical networks. The goal is a core set of middleware services that will allow scientists to set up secure, controlled environments for collaborative sharing of distributed resources for their research. Collectively, these middleware services and the global high-speed research networks will constitute the new Cyberinfrastructure (in the United States) or e-Infrastructure (in Europe) for collaborative scientific research.

The term "Grid" was first used in the mid-1990s to denote a distributed computing infrastructure for advanced science and engineering. At that time, the idea was driven by a desire to use distributed computing resources as a metacomputer, and the name was taken from the electricity power grid, with the analogy that computing power would be made available for anyone anywhere to use. The Grid was a product of developing technologies in high-performance computers and networking, together with the 1980s Grand Challenges research program in the United States. In 2001,

Ian Foster, Carl Kesselman, and Steve Tuecke recognized the broader relevance of the Grid and redefined the Grid in terms of infrastructure to facilitate collaboration (*7*).

Unfortunately, present-day versions of Grid middleware provide only a small part of the functionality required for e-Science collaborations. Nevertheless, the vision of a set of middleware services that will allow scientists to set up VOs tailored to the needs of their specific e-Science communities has proved to have universal appeal. This vision is at the heart of the UK's e-Science program (*8*) and a similar vision is embodied in the Atkins report on Cyberinfrastructure for NSF (*9*).

Web Service Grids

Web Services are the distributed computing technology that the IT industry is uniting around to be the building blocks for interoperable, distributed IT systems (*10*). By encapsulating internal resources within the service and providing a layer of application logic between those resources and the consumers, the owners of the service are free to evolve its internal structure over time (for example, to improve its performance or dependability), without making changes in the message exchange patterns used by existing service consumers. This encourages loose coupling between consumers and service providers, which is important for building robust inter-enterprise IT systems, because no one party is in complete control of all parts of the distributed application.

Web Services have largely been developed to build VOs in the private sector. Most of the Web Services standards are being done in the context of the World Wide Web Consortium (W3C). The scientific community has been extending Web Services for scientific applications in the context of the Global Grid Forum (GGF). It is developing an Open Grid Services Architecture (OGSA) based on Web Services (*11*, *12*). By leveraging developments in Web Services technologies, e-Science application developers will be able to exploit the tools, documentation, educational materials, and experience from the Web Services community when building their applications. The e-Science community can focus on building the higher-level services specific to the application domain, while responsibility for the design of the basic components of a reliable underlying infrastructure is left to the IT industry. The GGF will soon publish standards and protocols for information services, execution management, data access and integration, resource management, and security. These basic services together with standards for portal technology and visualization services will enable scientists to use generic middleware infrastructure services to build their application-specific VOs. This is the rationale for the UK e-Science ''plug-and-play composable services'' vision for Grid middleware.

e-Science and Semantics

The UK e-Science program has around 100 projects covering many areas of science, engineering, and medicine. In areas such as astronomy and earth science, global communities are coming together to define common standards for data and metadata to allow sharing and access to information (*13*, *14*). Other scientists are using high-performance simulations, computational steering, and remote visualization to advance the state of the art in their respective fields (*15*, *16*). In engineering, companies such as Rolls Royce and BAESystems are exploring how such e-Science technology can assist them in exploiting new distributed applications (*17*, *18*). In bioinformatics, researchers and pharmaceutical companies are attempting to use e-Science technologies to reduce data to information and information to knowledge (*19*, *20*). And in medical informatics, there are ambitious projects on digital mammography and electronic patient records (*21*, *22*). Rather than enumerate such examples in detail, we shall look at two projects that are attempting to combine conventional data and computing technologies with technologies from the Semantic Web community (*23*).

The myGrid e-Science project is researching high-level middleware to support personalized in silico experiments in biology (*19*). These in silico experiments use databases and computational analysis rather than laboratory investigations to test hypotheses. In myGrid, the emphasis is on data-intensive experiments that combine the use of applications and database queries. These bioinformatics experiments often involve many processes or services that need to be orchestrated. Workflow tools enable this orchestration and help the biologist to design, describe, and record complex experiments in terms with which they can interact and that can also interact with the workflows of other researchers. Intermediate workflows and data are kept, notes and thoughts recorded, and different experiments linked together to form a network of evidence, as is currently done in bench laboratory notebooks.

The computer scientists and biologists in the project have together developed a detailed set of scenarios for investigation of the genetics of Graves' disease, an immune disorder causing hyperthyroidism, and of Williams-Beuren syndrome, a gene deletion disorder that affects multiple human systems and also causes mental retardation. To implement its ideas, the project has built a prototype electronic workbench based on Web Services. They have identified four categories of service: (i) external third-party services, such as databases, computational analyses, and simulations, wrapped as Web Services; (ii) services for forming and executing experiments, such as workflows, information management, and distributed database query processing; (iii) services for supporting the e-Science methodology, such as provenance and notification; (iv) semantic services, such as service registries, ontologies, and ontology management, that enable the user to discover services and workflows and to manage several different types of metadata. Some or all of these services are then used to support applications and build application services.

The project has developed a suite of ontologies (roughly speaking, agreed-on vocabularies of terms or concepts) to represent metadata associated with the different middleware services. Semantic Web technologies such as DAML+OIL (*24*) and standards body W3C's Web ontology language OWL (*25*) then allow the prototype myGrid workbench to reason over these services intelligently. The project has demonstrated the potential of such an approach for in silico bioinformatics experiments and is now attempting to produce more robust semantic components that will allow users to personalize their own research environments (*26*–*28*).

Another such project, Comb*e*Chem, has the ambitious goal of creating a Smart Laboratory for chemistry, using technologies for automation, semantics, and Grid computing (*29*–*31*). A key driver of the project is the fact that large volumes of new chemical data are being created by new high-throughput technologies, such as combinatorial chemistry, in which large numbers of new chemical compounds are synthesized simultaneously. The need for assistance in organizing, annotating, and searching this data is becoming acute. The multidisciplinary Comb*e*Chem team have therefore developed a prototype Smart Laboratory test bed that integrates chemical structure and property data resources with a Grid-based computing environment. The project has explored automated procedures for finding similarities in solid-state crystal structures across families of compounds and has evaluated new statistical design concepts to improve the efficiency of combinatorial experiments in the search for new enzymes and pharmaceutical salts for improved drug delivery. One of the key concepts of the Comb*e*Chem project is Publication@Source, though which there is a complete end-to-end connection between the results obtained at the laboratory bench and the final published analyses (*32*). In a sister project called eBank, raw crystallographic data are annotated with metadata and published by being archived in the UK National Data Store as a crystallographic e-print (*33*). Publications can then be linked back to the raw data for other researchers to access. The project has a vision for what they call a scholarly cycle, encompassing experimentation, analysis, publication, research, and learning (Fig. 1).

In another strand, computer scientists in the SmartTea project have worked with the Comb*e*Chem team to develop an innovative

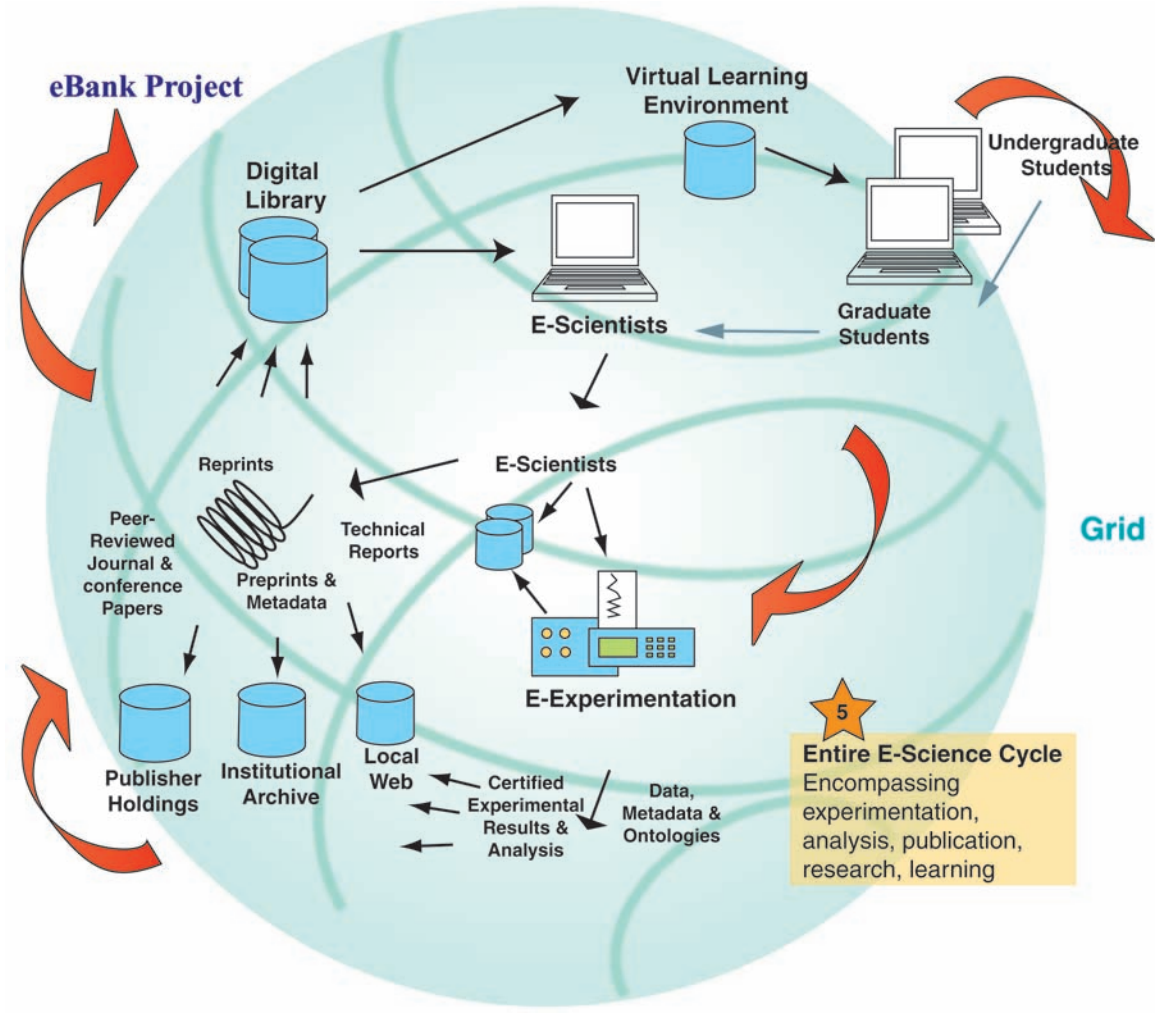


Fig. 1. The UK eBank project is focused on the changing landscape of scholarly communication, building links from e-research to e-learning, facilitating the scholarly knowledge cycle through the integration of digital repositories (experimental data, e-prints, and learning objects), and providing aggregator services. [Image courtesy of Liz Lyon and the eBank team]

human-centered system that captures the process of a chemistry experiment from plan to execution (*34*, *35*). They have used an analysis of the process of making tea in a laboratory to develop an electronic lab book replacement. Using tablet PCs, the system has undergone a successful trial in a synthetic organic chemistry laboratory and is linked to a flexible back-end storage system. A key finding was that users needed to feel in control, and this necessitated a high degree of flexibility in the lab book/user interface. The computer scientists on the team investigated the representation and storage of human-scale experiment metadata and introduced an ontology to describe the record of an experiment and a novel storage system for the data from the electronic lab book. In the same way in which the interfaces needed to be flexible to cope with whatever chemists wished to record, the back-end solutions needed to be flexible to store any metadata that might be created. Their storage system was based on Semantic Web technologies such as RDF (Resource Description Framework) and Web Services. This system was found to give a much higher degree of flexibility to the type of metadata that can be stored, as compared to traditional relational databases.

Toward a Semantic Grid

In 2001, De Roure, Jennings, and Shadbolt introduced the notion of the Semantic Grid, which advocated "the application of Semantic Web technologies both on and in the Grid" (*36*). From the requirements derived from the diverse set of UK e-Science applications, they identified a need for maximum reuse of software, services, information, and knowledge. Although the basic Grid middleware was originally conceived for hiding the heterogeneity of distributed computing, the authors contended that users now required "interoperability across time as well as space" to cope with both anticipated and unanticipated reuse of services, information, and knowledge. In a new paper, the same authors have revisited the e-Science program 3 years on from their original analysis to examine whether their expectations have been realized (*37*). They now see the e-Science requirements as a spectrum, with one end characterized by automation, virtual organizations of services, and the digital world, and the other end characterized by interaction, virtual organizations of people, and the physical world.

Conclusions

The broad view of Cyberinfrastructure/e-Infrastructure/Grid middleware services represented by the UK e-Science vision of plug-and-play composable middleware represents an exciting opportunity for both scientists and computer scientists. Although there is currently much focus in the Grid community on the low-level middleware, there are substantial research challenges for computer scientists to develop high-level intelligent middleware services that genuinely support the needs of scientists and allow them to routinely construct secure VOs and manage the veritable deluge of scientific data that will be generated in the next few years.

In the United Kingdom, we have therefore initiated a research program complementary to the e-Science application projects, whose goal is to explore the long-term computer science challenges arising from e-Science requirements (*38*). However, in parallel with this research thread, there is also the need to capture the prototype generic middleware services developed by our research projects and reengineer them for reuse by others. It is a major software engineering challenge to ensure that middleware components developed in the United Kingdom will interoperate with those developed in the United States, Asia, and elsewhere in Europe. This is the challenging mission for our newly established Open Middleware Infrastructure Institute (*39*).

In this article we have restricted our e-Science examples to those in the UK program (*40*). Needless to say, there are many other interesting e-Science projects in many countries of the world. Together, this global e-Science community is making progress toward realizing Licklider's vision for the Internet and in creating the components for a global middleware infrastructure. But there is still a long way to go before such middleware services can be used routinely by scientists going about their research.

References and Notes

1. S. Segaller, *Nerds: A Brief History of the Internet* (TV Books, New York, 1998), quote by L. Roberts.
2. T. Hey, A. E. Trefethen, in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, G. Fox, T. Hey, Eds. (Wiley, New York, 2003), pp. 809–824.
3. J. M. Taylor, see www.e-science.clrc.ac.uk.
4. I. J. R. Aitchison, A. J. G. Hey, *Gauge Theories in Particle Physics* (Institute of Physics, Bristol, UK, ed. 3, 2004).
5. The LHC Computing Grid (lhcgrid.web.cern.ch/LHCgrid).
6. The Integrative Biology Project (www.integrativebiology.ox.ac.uk).
7. I. Foster, C. Kesselman, S. Tuecke, *Int. J. Supercomputer Appl.* **15**, 3 (2001).
8. T. Hey, A. Trefethen, *Future Generation Comp. Syst.* **18**, 1017 (2002).
9. Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure, *Revolutionizing Science and Engineering Through Cyberinfrastructure*, D. Atkins *et al*., Eds. (www.nsf.gov).
10. W3C, *Web Services Architecture* (www.w3.org/TR/2004/NOTE-ws-arch-20040211) (2004).
11. For a definition of service-oriented architecture, see www.service-architecture.com/web-services/articles/service-oriented_architecture_soa_definition.html.
12. OGSA Working Group, GGF (www.ggf.org).
13. The International Virtual Observatory Alliance (www.ivoa.net).
14. The NERC DataGrid Project (ndg.nerc.ac.uk).
15. J. Chin *et al.*, *Contemp. Phys.* **44**, 417 (2003).
16. The GENIE Project (www.genie.ac.uk).
17. The DAME Project (www.cs.york.ac.uk/dame).
18. The GeWiTTS Project (www.nesc.ac.uk/events/sc2004/talks).
19. The myGrid Project (www.mygrid.org).
20. The e-Family Project (www.sanger.ac.uk/xml/eFamily.xsd).
21. The eDiaMoND Project (www.ediamond.ox.ac.uk).
22. The CLEF Project (www.clinical-escience.org).
23. T. Berners-Lee, J. Hendler, O. Lassila, *Sci. Am.* **284**, 34 (2001).
24. DAML+OIL (www.daml.org).
25. W3C OWL (www.w3.org/TR/owl-features).
26. R. Stevens *et al*., '*Exploring Williams-Beuren Syndrome using myGrid*', published in the Proceedings of the Conference on Intelligent Systems for Molecular Biology (ISMB), 2004.
27. C. Wroe *et al.*, *IEEE Intell. Syst.* **19**, 48 (2004).
28. S. Miles *et al.*, *IEEE Proc. Software* **150**, 252 (2003).
29. The CombeChem project (www.CombeChem.org).
30. J. G. Frey *et al*., in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, G. Fox, T. Hey, Eds. (Wiley, New York, 2003), pp. 945–962.
31. G. Hughes, H. Mills, D. De Roure, J. G. Frey, L. Moreau, *Org. Biomol. Chem.* **2**, 3284 (2004).
32. J. G. Frey, D. De Roure, L. A. Carr, *Publication at Source: Scientific Communication from a Publication Web to a Data Grid*, Euroweb 2002 Conference, The Web and the Grid: From e-Science to e-Business.
33. The SmartTea project (www.SmartTea.org).
34. m. c. schraefel *et al*., Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems, Vienna, Austria, 24 to 29 April 2004.
35. m. c. schraefel, G. Hughes, H. Mills, G. Smith, J. G. Frey, in *Proceedings of the Conference on Designing Interactive Systems* (Cambridge, MA, in press).
36. D. De Roure, N. R. Jennings, N. R. Shadbolt, *Research Agenda for the Semantic Grid: A Future e-Science Infrastructure* (National e-Science Centre, Edinburgh, UK, 2001).
37. D. De Roure, N. R. Jennings, N. R. Shadbolt, *Proc. IEEE* **93**, 669 (2005).
38. EPSRC Computer Science for e-Science Program (www.epsrc.ac.uk/ResearchFunding/Programmes/e-Science/ComputerScienceforeScience).
39. The Open Middleware Infrastructure Institute (www.omii.ac.uk).
40. The UK e-Science Programme (www.rcuk.ac.uk/escience).

41. The authors thank their colleagues in the e-Science Programme (S. Cox, D. De Roure, J. Frey, A. Keane, L. Moreau, N. Shadbolt, C. Goble, and D. Gavaghan) for many interesting and educational discussions, J. Gray for his insightful comments on the draft of this article, and L. Lyon and her team for assistance with the eBank image. The authors acknowledge the support of the UK e-Science Programme.

10.1126/science.1110410

VIEWPOINT

Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research

Kenneth H. Buetow

Biomedicine has experienced explosive growth, fueled in parts by the substantial increase of government support, continued development of the biotechnology industry, and the increasing adoption of molecular-based medicine. At its core, it is composed of fiercely independent, innovative, entrepreneurial individuals, organizations, and institutions. The field has developed unprecedented capacity to characterize biologic systems at their most fundamental levels with the use of tools and technologies almost unimaginable a generation ago. Biomedicine is at the precipice of unlocking the very essence of biologic life and enabling a new generation of medicine. Development and deployment of cyberinfrastructure may prove to be on the critical path to obtaining these goals.

The biomedical research community, dynamic and technology driven, shares its information through approaches initiated with Gutenberg's printing press and conceptually recognizable to scientists in the 18th century. Scientific findings are captured, summarized, and shared through manuscripts. The information infrastructure revolution that has transformed business and has had marked impact in other scientific disciplines has had slow uptake in biology and medicine.

Unquestionably, tremendous progress has been made in biomedicine through the application of information technology to this traditional information-sharing process. E-papers and e-journals and indices such as Pubmed all facilitate the sharing of manuscripts. Increasingly, biomedical journals require that primary data be deposited on a publisher's or investigator's Web-accessible site. In some communities, large centralized repository databases have been created for archiving biologic findings. These repositories support information retrieval through evolving current-art information technology [such as file transfer protocol (FTP) sites and Web browser portals]. For example, a recent plug-in for the Firefox Web browser permits researchers to have keyword access to these disparate data resources. However, like the communities that generate them, the infrastructure and information generated in biomedicine are largely disconnected and disjoint. Similarly, biomedical informatics, which I define as the application of information technology and its tools in biomedical disciplines (*1*), mirrors this structure of the culture it serves: highly heterogeneous in approach, small, independent, dispersed, and fragmented.

National Cancer Institute Center for Bioinformatics, National Institutes of Health, Rockville, MD 20892, USA. E-mail: buetowk@nih.gov

Biomedicine at a Crossroads

The current paradigms of information sharing and resource use in biology and medicine are being challenged on several fronts. First, the

In the United Kingdom, we have therefore initiated a research program complementary to the e-Science application projects, whose goal is to explore the long-term computer science challenges arising from e-Science requirements (*38*). However, in parallel with this research thread, there is also the need to capture the prototype generic middleware services developed by our research projects and reengineer them for reuse by others. It is a major software engineering challenge to ensure that middleware components developed in the United Kingdom will interoperate with those developed in the United States, Asia, and elsewhere in Europe. This is the challenging mission for our newly established Open Middleware Infrastructure Institute (*39*).

In this article we have restricted our e-Science examples to those in the UK program (*40*). Needless to say, there are many other interesting e-Science projects in many countries of the world. Together, this global e-Science community is making progress toward realizing Licklider's vision for the Internet and in creating the components for a global middleware infrastructure. But there is still a long way to go before such middleware services can be used routinely by scientists going about their research.

References and Notes

1. S. Segaller, *Nerds: A Brief History of the Internet* (TV Books, New York, 1998), quote by L. Roberts.
2. T. Hey, A. E. Trefethen, in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, G. Fox, T. Hey, Eds. (Wiley, New York, 2003), pp. 809–824.
3. J. M. Taylor, see www.e-science.clrc.ac.uk.
4. I. J. R. Aitchison, A. J. G. Hey, *Gauge Theories in Particle Physics* (Institute of Physics, Bristol, UK, ed. 3, 2004).
5. The LHC Computing Grid (lhcgrid.web.cern.ch/LHCgrid).
6. The Integrative Biology Project (www.integrativebiology.ox.ac.uk).
7. I. Foster, C. Kesselman, S. Tuecke, *Int. J. Supercomputer Appl.* **15**, 3 (2001).
8. T. Hey, A. Trefethen, *Future Generation Comp. Syst.* **18**, 1017 (2002).
9. Report of the National Science Foundation Blue-Ribbon Panel on Cyberinfrastructure, *Revolutionizing Science and Engineering Through Cyberinfrastructure*, D. Atkins *et al*., Eds. (www.nsf.gov).
10. W3C, *Web Services Architecture* (www.w3.org/TR/2004/NOTE-ws-arch-20040211) (2004).
11. For a definition of service-oriented architecture, see www.service-architecture.com/web-services/articles/service-oriented_architecture_soa_definition.html.
12. OGSA Working Group, GGF (www.ggf.org).
13. The International Virtual Observatory Alliance (www.ivoa.net).
14. The NERC DataGrid Project (ndg.nerc.ac.uk).
15. J. Chin *et al.*, *Contemp. Phys.* **44**, 417 (2003).
16. The GENIE Project (www.genie.ac.uk).
17. The DAME Project (www.cs.york.ac.uk/dame).
18. The GeWiTTS Project (www.nesc.ac.uk/events/sc2004/talks).
19. The myGrid Project (www.mygrid.org).
20. The e-Family Project (www.sanger.ac.uk/xml/eFamily.xsd).
21. The eDiaMoND Project (www.ediamond.ox.ac.uk).
22. The CLEF Project (www.clinical-escience.org).
23. T. Berners-Lee, J. Hendler, O. Lassila, *Sci. Am.* **284**, 34 (2001).
24. DAML+OIL (www.daml.org).
25. W3C OWL (www.w3.org/TR/owl-features).
26. R. Stevens *et al*., '*Exploring Williams-Beuren Syndrome using myGrid*', published in the Proceedings of the Conference on Intelligent Systems for Molecular Biology (ISMB), 2004.
27. C. Wroe *et al.*, *IEEE Intell. Syst.* **19**, 48 (2004).
28. S. Miles *et al.*, *IEEE Proc. Software* **150**, 252 (2003).
29. The CombeChem project (www.CombeChem.org).
30. J. G. Frey *et al*., in *Grid Computing: Making the Global Infrastructure a Reality*, F. Berman, G. Fox, T. Hey, Eds. (Wiley, New York, 2003), pp. 945–962.
31. G. Hughes, H. Mills, D. De Roure, J. G. Frey, L. Moreau, *Org. Biomol. Chem.* **2**, 3284 (2004).
32. J. G. Frey, D. De Roure, L. A. Carr, *Publication at Source: Scientific Communication from a Publication Web to a Data Grid*, Euroweb 2002 Conference, The Web and the Grid: From e-Science to e-Business.
33. The SmartTea project (www.SmartTea.org).
34. m. c. schraefel *et al*., Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems, Vienna, Austria, 24 to 29 April 2004.
35. m. c. schraefel, G. Hughes, H. Mills, G. Smith, J. G. Frey, in *Proceedings of the Conference on Designing Interactive Systems* (Cambridge, MA, in press).
36. D. De Roure, N. R. Jennings, N. R. Shadbolt, *Research Agenda for the Semantic Grid: A Future e-Science Infrastructure* (National e-Science Centre, Edinburgh, UK, 2001).
37. D. De Roure, N. R. Jennings, N. R. Shadbolt, *Proc. IEEE* **93**, 669 (2005).
38. EPSRC Computer Science for e-Science Program (www.epsrc.ac.uk/ResearchFunding/Programmes/e-Science/ComputerScienceforeScience).
39. The Open Middleware Infrastructure Institute (www.omii.ac.uk).
40. The UK e-Science Programme (www.rcuk.ac.uk/escience).

41. The authors thank their colleagues in the e-Science Programme (S. Cox, D. De Roure, J. Frey, A. Keane, L. Moreau, N. Shadbolt, C. Goble, and D. Gavaghan) for many interesting and educational discussions, J. Gray for his insightful comments on the draft of this article, and L. Lyon and her team for assistance with the eBank image. The authors acknowledge the support of the UK e-Science Programme.

10.1126/science.1110410

VIEWPOINT

Cyberinfrastructure: Empowering a "Third Way" in Biomedical Research

Kenneth H. Buetow

Biomedicine has experienced explosive growth, fueled in parts by the substantial increase of government support, continued development of the biotechnology industry, and the increasing adoption of molecular-based medicine. At its core, it is composed of fiercely independent, innovative, entrepreneurial individuals, organizations, and institutions. The field has developed unprecedented capacity to characterize biologic systems at their most fundamental levels with the use of tools and technologies almost unimaginable a generation ago. Biomedicine is at the precipice of unlocking the very essence of biologic life and enabling a new generation of medicine. Development and deployment of cyberinfrastructure may prove to be on the critical path to obtaining these goals.

The biomedical research community, dynamic and technology driven, shares its information through approaches initiated with Gutenberg's printing press and conceptually recognizable to scientists in the 18th century. Scientific findings are captured, summarized, and shared through manuscripts. The information infrastructure revolution that has transformed business and has had marked impact in other scientific disciplines has had slow uptake in biology and medicine.

Unquestionably, tremendous progress has been made in biomedicine through the application of information technology to this traditional information-sharing process. E-papers and e-journals and indices such as Pubmed all facilitate the sharing of manuscripts. Increasingly, biomedical journals require that primary data be deposited on a publisher's or investigator's Web-accessible site. In some communities, large centralized repository databases have been created for archiving biologic findings. These repositories support information retrieval through evolving current-art information technology [such as file transfer protocol (FTP) sites and Web browser portals]. For example, a recent plug-in for the Firefox Web browser permits researchers to have keyword access to these disparate data resources. However, like the communities that generate them, the infrastructure and information generated in biomedicine are largely disconnected and disjoint. Similarly, biomedical informatics, which I define as the application of information technology and its tools in biomedical disciplines (*1*), mirrors this structure of the culture it serves: highly heterogeneous in approach, small, independent, dispersed, and fragmented.

National Cancer Institute Center for Bioinformatics, National Institutes of Health, Rockville, MD 20892, USA. E-mail: buetowk@nih.gov

Biomedicine at a Crossroads

The current paradigms of information sharing and resource use in biology and medicine are being challenged on several fronts. First, the

success of the enterprise means that there has been a marked increase in the number of investigators, organizations, and institutions conducting biomedical research. Tracking the work and providing infrastructure to support the expansion are increasingly difficult. This expansion has resulted in a substantial number of new journals and Web sites. Although current information technology supports ready access, it does not address abstraction, integration, and interpretation of information. The diverse bioinformatics tools generated to consume and evaluate the data rarely interoperate. Commonly, the community demonstrates a willingness to share data and applications, but the number and diversity of components that must be assembled are overwhelming.

The very data generated in modern biomedicine presents a primary challenge to the researcher. Many of the new technologies used in today's research generate large volumes of rapidly expanding and ever-changing data. Although Moore's law and cheap disk space have reduced the impact of this growth, individual scientists and institutions are spending an increasing fraction of their effort and resources simply retrieving and processing data. Biologic data represents additional challenges. To integrate biologic data, one must traverse multiple orders of magnitude of scale and complexity. Ideally, in biology one would want to move seamlessly between biologic and chemical process, organelle, cell, organ, organ system, individual, family, community, and population. The diversity of data types that are explored in biomedicine is somewhat orthogonal. Technology permits the characterization of genomic, proteomic, metabolomic, image, and other large-scale characterizations.

All of the above is further confounded by the organization of biomedicine into research fields and disciplines. Such discipline focus generates an insidious challenge to information integration. Each community speaks its own scientific dialect. This community ''speciation'' results in reduced flow of information between disciplines, slowing the diffusion of knowledge and critical progress.

Finally, biomedicine's culture is at the nexus of a challenge faced by many other scientific fields: the need for ''big'' science and team science. The call for big science recognizes that many of the technology approaches required in biology and medicine are expensive, beyond the reach of individual investigators, and increasingly challenging the resource reserves of all but a few institutions. New paradigms are required to support these investigations. The push for team science also recognizes that many problems cross traditional discipline boundaries.

Cyberinfrastructure: A Third Way

A view that the current biomedical research culture is incompatible with team or big science is overly simplistic. It is clear that big science and team science will be necessary to achieve the goals of biology and medicine. However, the small, independent investigator is still the engine of innovative research. Widespread adoption of cyberinfrastructure represents an alternative in which the two approaches can be blended to create virtual team science. In so doing, the organization of biomedicine retains its entrepreneurial independent investigators whose insights and resources can be virtually joined through information technology. Big science contributes large-scale, raw material that feeds the virtual communities. Cyberinfrastructure empowers a reinvention of biomedicine without having to fundamentally change its basic culture or operational characteristics—a third way.

It is one thing to suggest that cyberinfrastructure could transform biomedicine and quite another thing to achieve this transformation. Fortunately, biomedicine can benefit from the long experience of other communities' embrace of informatics infrastructure to guide its approach. To address challenges in biomedicine, it must deliver in several key fronts. First, it must add perceivable value to the enterprise. In order to achieve widespread adoption, users must be motivated to do something different. Traditionally, this means they need to be able to do something they couldn't do without using the technology. Cyberinfrastructure shows great promise in this area because it has the ability to address the challenges of large, complex data sets. However, greater capacity may not be a sufficient driver, as demonstrated by current low penetration. Cyberinfrastructure will also need to enable new capabilities through the integration of communities and their disparate data types.

A primary lesson from other fields is that information technology has its greatest impact when it changes the way work can be performed. This may manifest itself through the apparent elimination of processing steps or the need to duplicate resources locally. Existing technologies permit the sharing and joining of common resources within virtual groups. However, the complex issues and diversity of biologic data still represent a substantial challenge to the creation of automated workflows.

Finally, the infrastructure needs to be easy to use and straightforward to implement. This requirement is more subtle than it might seem. A deeper examination raises the question, easy and straightforward to whom? Looking at the existing Internet and Web provides a useful clarification. End users consuming Internet resources through graphical user interfaces displayed through Web browsers would describe the Internet as easy to use. However, at the level of technical implementation, starting up a network that connects to the Internet and sharing information through a Web server is quite complex and beyond the skill set of an average biomedical researcher. It will be important to understand this dialectic as cyberinfrastructure is deployed across biology and medicine.

Biomedical research has experimented with the use of cyberinfrastructure to address the challenges outlined above for many years. An early example is found in the Cooperative Human Linkage Center (CHLC), a consortia formed early in the 1990s as part of the Human Genome Project for the purpose of creating genome-wide integrated genetic maps (*2*). CHLC was a geographically distributed virtual center connecting small specialized laboratories through informatics infrastructure communicating over the Internet (actually NSFnet at the time). It fulfilled a big-science need (creating the genetic map) through team science (each laboratory contributed specialized expertise) integrated virtually through current-art information technology. Each group worked in a context familiar to their specialized skills and the disparate parts were assembled by cyberinfrastructure to create the map. Map construction occurred through a pipelined workflow and used distributed processing over a network of multiuse computers. The raw data, analytic intermediates, and maps were distributed over the Internet through Web servers. The infrastructure to compute the maps was made available to the community through e-mail services. This example provides proof of concept that key aspects of the goals articulated above can be addressed, even with the use of a previous generation of information technology.

Technical Approach

The biology end user really doesn't care what technologies underlie cyberinfrastructure. Moreover, technology may not be the limiting factor in the development and deployment. However, the biomedical end user does provide key requirements that should be taken into consideration when choosing technology.

To facilitate adoption, cyberinfrastructure should be an extension of or interoperate with infrastructure already available to users. Ideally, it should integrate with and/or extend existing World Wide Web applications (supporting end-user needs) and Internet technology stacks (supporting the needs and existing investments of systems administrators where possible). Minimally, there must be a clear path from existing infrastructure to the new cyberinfrastructure.

The cyberinfrastructure vendor, operating system, and hardware should be as agnostic as possible. Users must have the capacity to change all of the above in order to maintain innovation and adjust to changing needs and developing technology. Open source is an oft-suggested solution to this. However, it can also be obtained by open standards and a commitment by those generating closed systems to adhere to these standards and to develop interfaces to communicate to and through them.

Biomedical cyberinfrastructure must also consider access and identity management as primary requirements. Although not unique to

biomedicine, protection of human subjects is required, as is the control and tracking of intellectual property and the need to establish academic credit and data provenance.

Many experiments are being implemented to explore alternative technologies that could possibly underlie cyberinfrastructure. These include peer-to-peer technology, Web services, and grid technology. Each has interesting potential. Grid technology has several distinguishing features (*3*, *4*). First, as a consequence of the widespread use of the Globus Toolkit (*5*) in various settings, grid technology is increasingly mature. Grid technology can support virtual communities through sharing of computational resources and data resources. Access and identity control are fundamental components of the architecture. The technology supports deterministic queries across a distributed, common schema. Its fundamental architecture also supports stateful processes important to the concept of workflow. The developing Open Grid Service Architecture–Data Access Integration (OGSA-DAI) framework holds promise for adding semantics to the grid technology so that computable, semantic interoperability may be achieved. Specific database schemas and data representations can be abstracted through a metadata layer. This information can be captured and shared in ontologies and services. This advance shows promise for machine capturing of information from the disparate biomedical communities and integrating of data and information into knowledge.

Grid architecture does have some key limitations. First, despite its developing research maturity, Grid is a distant second in commercial application. Web service architecture is the technology of choice for the vast majority of cyberinfrastructure support installations, in part because of the greater relative simplicity of the architecture. It is a straightforward extension of Internet and Web infrastructure familiar to the vast majority of systems designers and administrators. The broader developer and support base associated with Web services is important to the biomedical community.

Grid technology is not the only architecture with the capacity to address the challenges faced in biomedicine. However, what distinguishes Grid from, for example, Web services is that the capabilities described above are fundamental to the architecture. Web solutions to the challenges are outside the architecture and as such individually defined in each instance that they are created. The Grid architecture provides a standard framework for their representation and use. Encouragingly, Grid and Web services are converging.

Cyberinfrastructure in Action

As indicated above, the biomedical research community is conducting numerous experiments in developing and deploying cyberinfrastructure. With respect to Grid architecture, many are accessible through an index maintained by the Global Grid Forum (www.gridforum.org).

Many of these demonstration test beds explore the traditional definition of Grid computing in biomedicine, namely the sharing of resources across a virtual community. The range of these applications is impressive. They include molecular docking, protein structure determination, nucleic acid sequence alignment, and biologic feature extraction.

Several "proof-of-concept" test beds are exploring broad aspects of cyberinfrastructure in biomedicine, including the following:

Biomedical Informatics Research Network (BIRN). The BIRN project (www.nbirn.net) has focused on creation of geographically distributed virtual communities through shared resources. Its early work has been addressing the problems associated with new imaging platforms and the need to cross-correlate functional and structural data generated by these platforms. Its challenge is at the heart of cyberinfrastructure: How does one store, manage, curate, access, visualize, and analyze large volumes of data across a virtual community? Imaging projects generate terabytes of data through the use of disparate imaging technologies, all requiring compute-intensive applications to process.

BIRN has approached this problem by creating the virtual community through the distribution of a common, homogeneous, centrally configured hardware rack. This rack comes installed with appropriate software necessary to create the virtual community. The community is connected at high speed through the use of the Internet 2/Abiline backbone. It uses the Grid architecture defined by the Globus toolkit with numerous extensions, particularly in the areas of brokering storage resources across the community and the use of a metadata catalog.

A series of defined test beds are evaluating and extending the cyberinfrastructure, with a key focus of neuroimaging. Each test bed of defined members is exploring a dimension of the neuroimaging domain, with one centered around brain morphology, another around functional imaging (in schizophrenia), and the last around multiscale models in experimental systems (mouse).

myGrid. The myGrid project (www.mygrid.org.uk) takes a different perspective on application of cyberinfrastructure. Its focus is the support of investigator-driven experiments in silico. In myGrid, local and public data can be computationally evaluated to ask and answer questions in biology. It is less focused on resource sharing than BIRN, but rather strives to address issues related to semantic complexity of biologic data and the applications that process that data. It has constructed services that facilitate integration of data and applications. It addresses challenges associated with

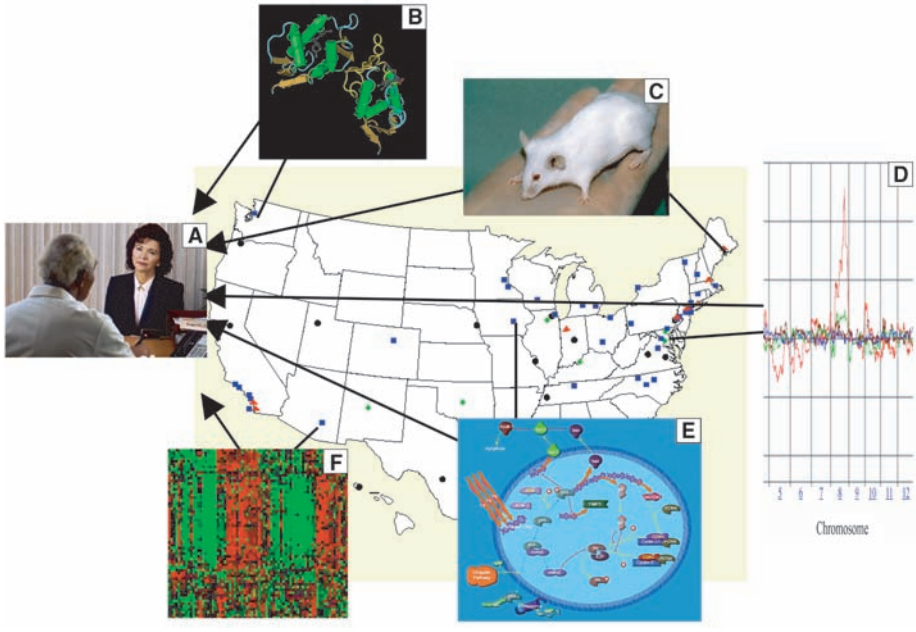


Fig. 1. The caBIG aims to integrate diverse biomedical research data so that investigators can consume data, services, and knowledge distributed throughout the research enterprise. For example, a scientist in California (**A**) designs an investigation following a computer modeling hypothesis-generating experiment where agent information from Washington (**B**) is queried in the context of animal model information from Maine (**C**). Genomic aspects of the experiment use comparative genome hybridization findings generated by colleagues in Maryland (**D**), which are interpreted in biologic processes from pathway data curated in Iowa (**E**). These are contrasted to reference expression signatures generated by researchers in Arizona (**F**).

the rapidly evolving nature of biomedical data and issue of data provenance. Particularly interesting is its approach to creating workflows. Within its framework it supports resource discovery and distributed queries.

myGrid is a service-based architecture whose core is Web services and OGSA-DAI. It uses the common Internet and does not require specialized hardware. It accomplishes its semantic interoperability through the use of ontology-based metadata. These metadata describe data, services, and other components of the infrastructure. The environment is open; however, it has the capacity to address the mixed data and service access requirements of researchers.

The myGrid project is exploring the diversity of the domains associated with biomedical cyberstructure. In one test bed, it has explored the circadian rhythms in *Drosophila melanogaster*. In a complementary test bed, it has supported genetic investigations of the human immune disorder Graves disease.

The cancer Biomedical Informatics Grid (caBIG). The approach of the caBIG project (http://caBIG.nci.nih.gov) to cyberinfrastructure is a conceptual hybrid between BIRN and myGrid. Similar to BIRN, its focus is to create a virtual community that shares resources and tackles the key issues of cyberinfrastructure. However, this community is open, spans the vast domain of cancer research, and is attempting to integrate the bench-to-bedside research cycle.

Similar to myGrid, it is an open infrastructure striving to achieve computational semantic interoperability. The caBIG's cyberinfrastructure is also a service-based architecture whose core is Web services and OGSA-DAI. It uses the common Internet and does not require specialized hardware. It has constructed services that facilitate integration of data and applications. Within its framework it supports resource discovery and distributed queries.

A key difference between myGrid and caBIG is the way they approach semantics and their related services. The caBIG cyberinfrastructure uses a common set of services and service registrations for the entire community. The shared caBIG semantic services provide biomedical ontologies and vocabularies in common use across biomedicine and cross-mappings between them. These mappings facilitate cross-disciplinary data integration and interpretation. The shared caBIG semantic services additionally include common data elements and object-based abstractions of the various research domains they serve. An open community process is used to maintain and extend these semantic resources. The use and registration of this common model-driven architecture serves as the basis of community-wide service descriptions. The caBIG test bed currently supports basic and translational research, clinical trials research, and tissue banking and pathology (Fig. 1). Participation in these groups is open.

Biomedical Cyberstructure and the Future

The above efforts suggest that it is technically feasible to knit the vibrant threads of biomedicine into a rich tapestry. There are still many challenges ahead, both technical and cultural. The differences indicate that there is not a single path joining biomedicine. As each effort reaches maturity it will be important to compare and contrast the lessons learned from their overlapping approaches. For example, how can the community ensure that existing individual, domain, and institution silos are not simply replaced with cybersilos?

Also, although these efforts are provocative, they have not yet crossed the threshold of demonstrated value. Evidence suggests that those in the field of biomedicine are receptive to exploring these alternatives but are still skeptical. Cyberinfrastructure appears to be up to the challenges confronting biomedical research. These are early but exciting times.

References

1. T. Hey, A. E. Trefethen, *Science* **308**, 817 (2005).
2. J. C. Murray *et al.*, *Science* **265**, 2049 (1994).
3. I. Foster, *Sci. Am.* **288**, 78 (April 2003).
4. I. Foster, *Science* **308**, 814 (2005).
5. I. Foster, C. Kesselman, *Int. J. Supercomput. Appl.* **11**, 115 (1997).

10.1126/science.1112120