Beyond the Tsunami: Developing the Infrastructure to Deal with Life Sciences Data

CHRISTOPHER SOUTHAN
GRAHAM CAMERON
EMBL-European Bioinformatics Institute

SCIENTIFIC REVOLUTIONS ARE DIFFICULT TO QUANTIFY, but the rate of data generation in science has increased so profoundly that we can simply examine a single area of the life sciences to appreciate the magnitude of this effect across all of them. Figure 1 on the next page tracks the dramatic increase in the number of individual bases submitted to the European Molecular Biology Laboratory Nucleotide Sequence Database[1] (EMBL-Bank) by the global experimental community. This submission rate is currently growing at 200% per annum.

Custodianship of the data is held by the International Nucleotide Sequence Database Collaboration (INSDC), which consists of the DNA Data Bank of Japan (DDBJ), GenBank in the U.S., and EMBL-Bank in the UK. These three repositories exchange new data on a daily basis. As of May 2009, the totals stood at approximately 250 billion bases in 160 million entries.

A recent submission to EMBL-Bank, accession number FJ982430, illustrates the speed of data generation and the effectiveness of the global bioinformatics infrastructure in responding to a health crisis. It includes the complete H1 subunit sequence of 1,699 bases from the first case of novel H1N1 influenza virus in Denmark. This was submitted on May 4, 2009, within days of

[1] www.ebi.ac.uk/embl

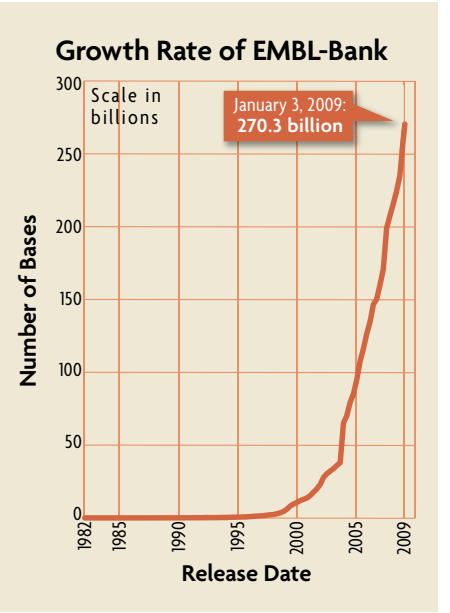


FIGURE 1.
Growth in the number of bases deposited in EMBL-Bank from 1982 to the beginning of 2009.

the infected person being diagnosed. Many more virus subunit sequences have been submitted from the U.S., Italy, Mexico, Canada, Denmark, and Israel since the beginning of the 2009 global H1N1 pandemic.

EMBL-Bank is hosted at the European Bioinformatics Institute (EBI), an academic organization based in Cambridge, UK, that forms part of the European Molecular Biology Laboratory (EMBL). The EBI is a center for both research and services in bioinformatics. It hosts biological data, including nucleic acid, protein sequences, and macromolecular structures. The neighboring Wellcome Trust Sanger Institute generates about 8 percent of the world's sequencing data output. Both of these institutions on the Wellcome Trust Genome campus include scientists who generate data and administer the databases into which it flows, biocurators who provide annotations, bioinformaticians who develop analytical tools, and research groups that seek biological insights and consolidate them through further experimentation. Consequently, it is a community in which issues surrounding computing infrastructure, data storage, and mining are confronted on a daily basis, and in which both local and global collaborative solutions are continually explored.

The collective name for the nucleotide sequencing information service is the European Nucleotide Archive [1]. It includes EMBL-Bank and three other repositories that were set up for new types of data generation: the Trace Archive for trace data from first-generation capillary instruments, the Short Read Archive for data from next-generation sequencing instruments, and a pilot Trace Assembly Archive that stores alignments of sequencing reads with links to finished genomic sequences in EMBL-Bank. Data from all archives are exchanged regularly with the National Center for Biotechnology Information in the U.S. Figure 2 compares the sizes of

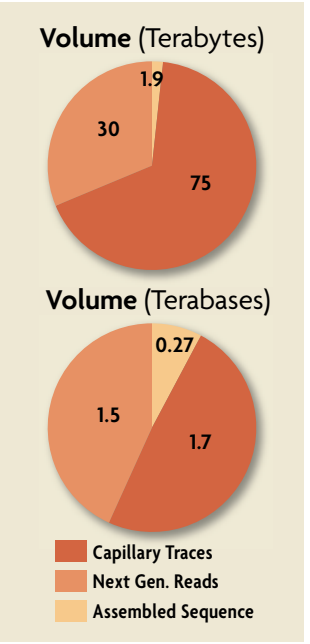


FIGURE 2.
The size in data volume and nucleotide numbers of EMBL-Bank, the Trace Archive, and the Short Read Archive as of May 2009.

EMBL-Bank, the Trace Archive, and the Short Read Archive.

THE CHALLENGE OF NEXT-GENERATION SEQUENCING

The introduction in 2005 of so-called next-generation sequencing instruments that are capable of producing millions of DNA sequence reads in a single run has not only led to a huge increase in genetic information but has also placed bioinformatics, and life sciences research in general, at the leading of edge of infrastructure development for the storage, movement, analysis, interpretation, and visualization of petabyte-scale datasets [2]. The Short Read Archive, the European repository for accepting data from these machines, received 30 terabytes (TB) of data in the first six months of operation—equivalent to almost 30% of the entire EMBL-Bank content accumulated over the 28 years since data collection began. The uptake of new instruments and technical developments will not only increase submissions to this archive manyfold within a few years, but it will also prelude the arrival of "next-next-generation" DNA sequencing systems [3].

To meet this demand, the EBI has increased storage from 2,500 TB (2.5 PB) in 2008 to 5,000 TB (5 PB) in 2009—an approximate annual doubling. Even if the capacity keeps pace, bottlenecks might emerge as I/O limitations move to other points in the infrastructure. For example, at this scale, traditional backup becomes impractically slow. Indeed, a hypothetical total data loss at the EBI is estimated to require months of restore time. This means that streamed replication of the original data is becoming a more efficient option, with copies being stored at multiple locations. Another bottleneck example is that technical advances in data transfer speeds now exceed the capacity to write out to disks—about 70 megabits/sec, with no imminent expectation of major performance increases. The problem can be ameliorated by writing to multiple disks, but at a considerable increase in cost.

This inexorable load increase necessitates continual assessment of the balance

between submitting derived data to the repositories and storing raw instrument output locally. Scientists at all stages of the process, experimentalists, instrument operators, datacenter administrators, bioinformaticians, and biologists who analyze the results will need to be involved in decisions about storage strategies. For example, in laboratories running high-throughput sequencing instruments, the cost of storing raw data for a particular experiment is already approaching that of repeating the experiment. Researchers may balk at the idea of deleting raw data after processing, but this is a pragmatic option that has to be considered. Less controversial solutions involve a triage of data reduction options between raw output, base calls, sequence reads, assemblies, and genome consensus sequences. An example of such a solution is FASTQ, a text-based format for storing both a nucleotide sequence and its corresponding quality scores, both encoded with a single ASCII character. Developed by the Sanger Institute, it has recently become a standard for storing the output of next-generation sequencing instruments. It can produce a 200-fold reduction in data volume—that is, 99.5% of the raw data can be discarded. Even more compressed sequence data representations are in development.

GENOMES: ROLLING OFF THE PRODUCTION LINE

The production of complete genomes is rapidly advancing our understanding of biology and evolution. The impressive progress is illustrated in Figure 3, which depicts the increase of genome sequencing projects in the Genomes OnLine Database (GOLD).

While the figure was generated based on all global sequencing projects, many of these genomes are available for analysis on the Ensembl Web site hosted jointly by the EBI and the Sanger Institute. The graph shows that, by 2010, well over 5,000 genome projects will have been initiated and more than 1,000 will have produced complete assemblies. A recent significant example is the bovine genome [4], which followed the chicken and will soon be joined by all other major agricultural species. These will not only help advance our understanding of mammalian evolution and domestication, but they will also accelerate genetic improvements for farming and food production.

RESEQUENCING THE HUMAN GENOME: ANOTHER DATA SCALE-UP

Recent genome-wide studies of human genetic variation have advanced our understanding of common human diseases. This has motivated the formation of an international consortium to develop a comprehensive catalogue of sequence variants in

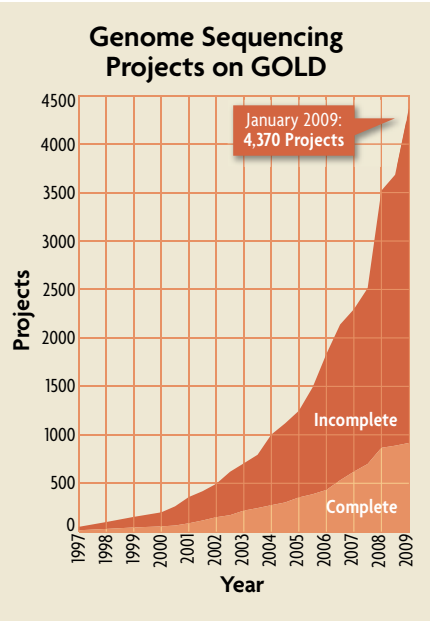


FIGURE 3.
The increase in both initiated and completed genome projects since 1997 in the Genomes OnLine Database (GOLD). Courtesy of GOLD.

multiple human populations. Over the next three years, the Sanger Institute, BGI Shenzhen in China, and the National Human Genome Research Institute's Large-Scale Genome Sequencing Program in the U.S. are planning to sequence a minimum of 1,000 human genomes.

In 2008, the pilot phase of the project generated approximately 1 terabase (trillion bases) of sequence data per month; the number is expected to double in 2009. The total generated will be about 20 terabases. The requirement of about 30 bytes of disk storage per base of sequence can be extrapolated to about 500 TB of data for the entire project. By comparison, the original human genome project took about 10 years to generate about 40 gigabases (billion bases) of DNA sequence. Over the next two years, up to 10 billion bases will be sequenced per day, equating to more than two human genomes (at 2.85 billion per human) every 24 hours. The completed dataset of 6 trillion DNA bases will be 60 times more sequence data than that shown earlier in Figure 1.

THE RAISON D'ÊTRE OF MANAGING DATA: CONVERSION TO NEW KNOWLEDGE

Even before the arrival of the draft human genome in 2001, biological databases were moving from the periphery to the center of modern life sciences research, leading to the problem that the capacity to mine data has fallen behind our ability to generate it. As a result, there is a pressing need for new methods to fully exploit not only genomic data but also other high-throughput result sets deposited in databases. These result sets are also becoming more hypothesis-neutral compared with traditional small-scale, focused experiments. Usage statistics for EBI services, shown in Figure 4 on the next page, show that the biological community, sup-

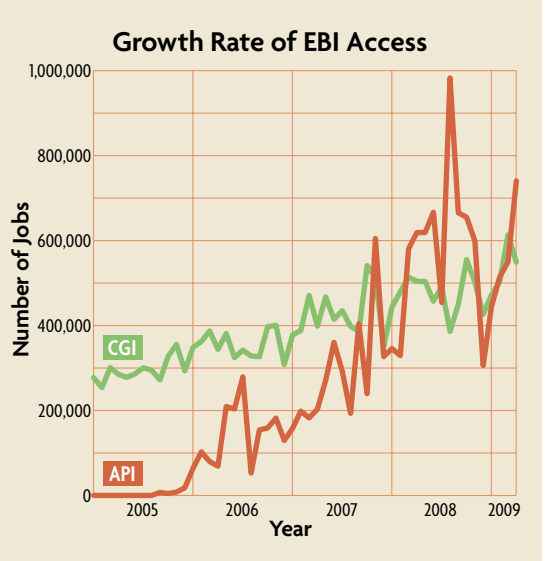


FIGURE 4.
Web accesses (Common Gateway Interface [CGI]) and Web services usage (application programming interface [API]) recorded on EBI servers from 2005 to 2009.

ported by the bioinformatics specialists they collaborate with, are accessing these resources in increasing numbers.

The Web pages associated with the 63 databases hosted at the EBI now receive over 3.5 million hits per day, representing more than half a million independent users per month. While this does not match the increase in rates of data accumulation, evidence for a strong increase in data mining is provided by the Web services' programmatic access figures, which are approaching 1 million jobs per month. To further facilitate data use, the EBI is developing, using open standards, the EB-eye search system to provide a single entry point. By indexing in various formats (e.g., flat files, XML dumps, and OBO format), the system provides fast access and allows the user to search globally across all EBI databases or individually in selected resources.

EUROPEAN PLANS FOR CONSOLIDATING INFRASTRUCTURE

EBI resources are effectively responding to increasing demand from both the generators and users of data, but increases in scale for the life sciences across the whole of Europe require long-term planning. This is the mission of the ELIXIR project, which aims to ensure a reliable distributed infrastructure to maximize access to biological information that is currently distributed in more than 500 databases throughout Europe. The project addresses not only data management problems but also sustainable funding to maintain the data collections and global collaborations. It is also expected to put in place processes for developing collections for new data

types, supporting interoperability of bioinformatics tools, and developing bioinformatics standards and ontologies.

The development of ELIXIR parallels the transition to a new phase in which high-performance, data-intensive computing is becoming essential to progress in the life sciences [5]. By definition, the consequences for research cannot be predicted with certainty. However, some pointers can be given. By mining not only the increasingly comprehensive datasets generated by genome sequencing mentioned above but also transcript data, proteomics information, and structural genomics output, biologists will obtain new insights into the processes of life and their evolution. This will in turn facilitate new predictive power for synthetic biology and systems biology. Beyond its profound impact on the future of academic research, this data-driven progress will also translate to the more applied areas of science—such as pharmaceutical research, biotechnology, medicine, public health, agriculture, and environmental science—to improve the quality of life for everyone.

REFERENCES

[1] G. Cochrane et al., "Petabyte-scale innovations at the European Nucleotide Archive," *Nucleic Acids Res.*, vol. 37, pp. D19–25, Jan. 2009, doi: 10.1093/nar/gkn765.

[2] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends Genet.*, vol. 24, no. 3, pp. 133–141, Mar. 2008, doi: 10.1016/j.physletb.2003.10.071.

[3] N. Blow, "DNA sequencing: generation next-next," *Nat. Methods*, vol. 5, pp. 267–274, 2008, doi: 10.1038/nmeth0308-267.

[4] Bovine Genome Sequencing and Analysis Consortium, "The genome sequence of taurine cattle: a window to ruminant biology and evolution," *Science*, vol. 324, no. 5926, pp. 522–528, Apr. 24, 2009, doi: 10.1145/1327452.1327492.

[5] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, Mar. 6, 2009, doi: 10.1126/science.1170411.

Multicore Computing and Scientific Discovery

JAMES LARUS
DENNIS GANNON
Microsoft Research

In the past half century, parallel computers, parallel computation, and scientific research have grown up together. Scientists and researchers' insatiable need to perform more and larger computations has long exceeded the capabilities of conventional computers. The only approach that has met this need is parallelism—computing more than one operation simultaneously. At one level, parallelism is simple and easy to put into practice. Building a parallel computer by replicating key operating components such as the arithmetic units or even complete processors is not difficult. But it is far more challenging to build a well-balanced machine that is not stymied by internal bottlenecks. In the end, the principal problem has been software, not hardware. Parallel programs are far more difficult to design, write, debug, and tune than sequential software—which itself is still not a mature, reproducible artifact.

THE EVOLUTION OF PARALLEL COMPUTING

The evolution of successive generations of parallel computing hardware has also forced a constant rethinking of parallel algorithms and software. Early machines such as the IBM Stretch, the Cray I, and the Control Data Cyber series all exposed parallelism as vector operations. The Cray II, Encore, Alliant, and many generations of IBM machines were built with multiple processors that

shared memory. Because it proved so difficult to increase the number of processors while sharing a single memory, designs evolved further into systems in which no memory was shared and processors shared information by passing messages. Beowulf clusters, consisting of racks of standard PCs connected by Ethernet, emerged as an economical approach to supercomputing. Networks improved in latency and bandwidth, and this form of distributed computing now dominates supercomputers. Other systems, such as the Cray multi-threaded platforms, demonstrated that there were different approaches to addressing shared-memory parallelism. While the scientific computing community has struggled with programming each generation of these exotic machines, the mainstream computing world has been totally satisfied with sequential programming on machines where any parallelism is hidden from the programmer deep in the hardware.

In the past few years, parallel computers have entered mainstream computing with the advent of multicore computers. Previously, most computers were sequential and performed a single operation per time step. Moore's Law drove the improvements in semiconductor technology that doubled the transistors on a chip every two years, which increased the clock speed of computers at a similar rate and also allowed for more sophisticated computer implementations. As a result, computer performance grew at roughly 40% per year from the 1970s, a rate that satisfied most software developers and computer users. This steady improvement ended because increased clock speeds require more power, and at approximately 3 GHz, chips reached the limit of economical cooling. Computer chip manufacturers, such as Intel, AMD, IBM, and Sun, shifted to multicore processors that used each Moore's Law generation of transistors to double the number of independent processors on a chip. Each processor ran no faster than its predecessor, and sometimes even slightly slower, but in aggregate, a multicore processor could perform twice the amount of computation as its predecessor.

PARALLEL PROGRAMMING CHALLENGES

This new computer generation rests on the same problematic foundation of software that the scientific community struggled with in its long experience with parallel computers. Most existing general-purpose software is written for sequential computers and will not run any faster on a multicore computer. Exploiting the potential of these machines requires new, parallel software that can break a task into multiple pieces, solve them more or less independently, and assemble the results into a single answer. Finding better ways to produce parallel software is currently

the most pressing problem facing the software development community and is the subject of considerable research and development.

The scientific and engineering communities can both benefit from these urgent efforts and can help inform them. Many parallel programming techniques originated in the scientific community, whose experience has influenced the search for new approaches to programming multicore computers. Future improvements in our ability to program multicore computers will benefit all software developers as the distinction between the leading-edge scientific community and general-purpose computing is erased by the inevitability of parallel computing as the fundamental programming paradigm.

One key problem in parallel programming today is that most of it is conducted at a very low level of abstraction. Programmers must break their code into components that run on specific processors and communicate by writing into shared memory locations or exchanging messages. In many ways, this state of affairs is similar to the early days of computing, when programs were written in assembly languages for a specific computer and had to be rewritten to run on a different machine. In both situations, the problem was not just the lack of reusability of programs, but also that assembly language development was less productive and more error prone than writing programs in higher-level languages.

ADDRESSING THE CHALLENGES

Several lines of research are attempting to raise the level at which parallel programs can be written. The oldest and best-established idea is data parallel programming. In this programming paradigm, an operation or sequence of operations is applied simultaneously to all items in a collection of data. The granularity of the operation can range from adding two numbers in a data parallel addition of two matrices to complex data mining calculations in a map-reduce style computation [1]. The appeal of data parallel computation is that parallelism is mostly hidden from the programmer. Each computation proceeds in isolation from the concurrent computations on other data, and the code specifying the computation is sequential. The developer need not worry about the details of moving data and running computations because they are the responsibility of the runtime system. GPUs (graphics processing units) provide hardware support for this style of programming, and they have recently been extended into GPGPUs (general-purpose GPUs) that perform very high-performance numeric computations.

Unfortunately, data parallelism is not a programming model that works for all

types of problems. Some computations require more communication and coordination. For example, protein folding calculates the forces on all atoms in parallel, but local interactions are computed in a manner different from remote interactions. Other examples of computations that are hard to write as data parallel programs include various forms of adaptive mesh refinement that are used in many modern physics simulations in which local structures, such as clumps of matter or cracks in a material structure, need finer spatial resolution than the rest of the system.

A new idea that has recently attracted considerable research attention is transactional memory (TM), a mechanism for coordinating the sharing of data in a multicore computer. Data sharing is a rich source of programming errors because the developer needs to ensure that a processor that changes the value of data has exclusive access to it. If another processor also tries to access the data, one of the two updates can be lost, and if a processor reads the data too early, it might see an inconsistent value. The most common mechanism for preventing this type of error is a lock, which a program uses to prevent more than one processor from accessing a memory location simultaneously. Locks, unfortunately, are low-level mechanisms that are easily and frequently misused in ways that both allow concurrent access and cause deadlocks that freeze program execution.

TM is a higher-level abstraction that allows the developer to identify a group of program statements that should execute atomically—that is, as if no other part of the program is executing at the same time. So instead of having to acquire locks for all the data that the statements might access, the developer shifts the burden to the runtime system and hardware. TM is a promising idea, but many engineering challenges still stand in the way of its widespread use. Currently, TM is expensive to implement without support in the processors, and its usability and utility in large, real-world codes is as yet undemonstrated. If these issues can be resolved, TM promises to make many aspects of multicore programming far easier and less error prone.

Another new idea is the use of functional programming languages. These languages embody a style of programming that mostly prohibits updates to program state. In other words, in these languages a variable can be given an initial value, but that value cannot be changed. Instead, a new variable is created with the new value. This style of programming is well suited to parallel programming because it eliminates the updates that require synchronization between two processors. Parallel, functional programs generally use mutable state only for communication among parallel processors, and they require locks or TM only for this small, distinct part of their data.

Until recently, only the scientific and engineering communities have struggled with the difficulty of using parallel computers for anything other than the most embarrassingly parallel tasks. The advent of multicore processors has changed this situation and has turned parallel programming into a major challenge for all software developers. The new ideas and programming tools developed for mainstream programs will likely also benefit the technical community and provide it with new means to take better advantage of the continually increasing power of multicore processors.

REFERENCES

[1] D. Gannon and D. Reed, "Parallelism and the Cloud," in this volume.

Parallelism and the Cloud

DENNIS GANNON
DAN REED
Microsoft Research

Over the past decade, scientific and engineering research via computing has emerged as the third pillar of the scientific process, complementing theory and experiment. Several national studies have highlighted the importance of computational science as a critical enabler of scientific discovery and national competitiveness in the physical and biological sciences, medicine and healthcare, and design and manufacturing [1-3].

As the term suggests, computational science has historically focused on computation: the creation and execution of mathematical models of natural and artificial processes. Driven by opportunity and necessity, computational science is expanding to encompass both computing and data analysis. Today, a rising tsunami of data threatens to overwhelm us, consuming our attention by its very volume and diversity. Driven by inexpensive, seemingly ubiquitous sensors, broadband networks, and high-capacity storage systems, the tsunami encompasses data from sensors that monitor our planet from deep in the ocean, from land instruments, and from space-based imaging systems. It also includes environmental measurements and healthcare data that quantify biological processes and the effects of surrounding conditions. Simply put, we are moving from data paucity to a data plethora, which is leading to a relative poverty of human attention to any individual datum

and is necessitating machine-assisted winnowing.

This ready availability of diverse data is shifting scientific approaches from the traditional, hypothesis-driven scientific method to science based on exploration. Researchers no longer simply ask, "What experiment could I construct to test this hypothesis?" Increasingly, they ask, "What correlations can I glean from extant data?" More tellingly, one wishes to ask, "What insights could I glean if I could fuse data from multiple disciplines and domains?" The challenge is analyzing many petabytes of data on a time scale that is practical in human terms.

The ability to create rich, detailed models of natural and artificial phenomena and to process large volumes of experimental data created by a new generation of scientific instruments that are themselves powered by computing makes computing a universal intellectual amplifier, advancing all of science and engineering and powering the knowledge economy. Cloud computing is the latest technological evolution of computational science, allowing groups to host, process, and analyze large volumes of multidisciplinary data. Consolidating computing and storage in very large datacenters creates economies of scale in facility design and construction, equipment acquisition, and operations and maintenance that are not possible when these elements are distributed. Moreover, consolidation and hosting mitigate many of the sociological and technical barriers that have limited multidisciplinary data sharing and collaboration. Finally, cloud hosting facilitates long-term data preservation—a task that is particularly challenging for universities and government agencies and is critical to our ability to conduct longitudinal experiments.

It is not unreasonable to say that modern datacenters and modern supercomputers are like twins separated at birth. Both are massively parallel in design, and both are organized as a network of communicating computational nodes. The individual nodes of each are based on commodity microprocessors that have multiple cores, large memories, and local disk storage. They both execute applications that are designed to exploit massive amounts of parallelism. Their differences lie in their evolution. Massively parallel supercomputers have been designed to support computation with occasional bursts of input/output and to complete a single massive calculation as fast as possible, one job at a time. In contrast, datacenters direct their power outward to the world and consume vast quantities of input data.

Parallelism can be exploited in cloud computing in two ways. The first is for human access. Cloud applications are designed to be accessed as Web services, so they are organized as two or more layers of processes. One layer provides the service interface to the user's browser or client application. This "Web role" layer accepts us-

ers' requests and manages the tasks assigned to the second layer. The second layer of processes, sometimes known as the "worker role" layer, executes the analytical tasks required to satisfy user requests. One Web role and one worker role may be sufficient for a few simultaneous users, but if a cloud application is to be widely used—such as for search, customized maps, social networks, weather services, travel data, or online auctions—it must support thousands of concurrent users.

The second way in which parallelism is exploited involves the nature of the data analysis tasks undertaken by the application. In many large data analysis scenarios, it is not practical to use a single processor or task to scan a massive dataset or data stream to look for a pattern—the overhead and delay are too great. In these cases, one can partition the data across large numbers of processors, each of which can analyze a subset of the data. The results of each "sub-scan" are then combined and returned to the user.

This "map-reduce" pattern is frequently used in datacenter applications and is one in a broad family of parallel data analysis queries used in cloud computing. Web search is the canonical example of this two-phase model. It involves constructing a searchable keyword index of the Web's contents, which entails creating a copy of the Web and sorting the contents via a sequence of map-reduce steps. Three key technologies support this model of parallelism: Google has an internal version [4], Yahoo! has an open source version known as Hadoop, and Microsoft has a map-reduce tool known as DryadLINQ [5]. Dryad is a mechanism to support the execution of distributed collections of tasks that can be configured into an arbitrary directed acyclic graph (DAG). The Language Integrated Query (LINQ) extension to C# allows SQL-like query expressions to be embedded directly in regular programs. The DryadLINQ system can automatically compile these queries into Dryad DAG, which can be executed automatically in the cloud.

Microsoft Windows Azure supports a combination of multi-user scaling and data analysis parallelism. In Azure, applications are designed as stateless "roles" that fetch tasks from queues, execute them, and place new tasks or data into other queues. Map-reduce computations in Azure consist of two pools of worker roles: mappers, which take map tasks off a map queue and push data to the Azure storage, and reducers, which look for reduce tasks that point to data in the storage system that need reducing. Whereas DryadLINQ executes a static DAG, Azure can execute an implicit DAG in which nodes correspond to roles and links correspond to messages in queues. Azure computations can also represent the parallelism generated by very large numbers of concurrent users.

This same type of map-reduce data analysis appears repeatedly in large-scale scientific analyses. For example, consider the task of matching a DNA sample against the thousands of known DNA sequences. This kind of search is an "embarrassingly parallel" task that can easily be sped up if it is partitioned into many independent search tasks over subsets of the data. Similarly, consider the task of searching for patterns in medical data, such as to find anomalies in fMRI scans of brain images, or the task of searching for potential weather anomalies in streams of events from radars.

Finally, another place where parallelism can be exploited in the datacenter is at the hardware level of an individual node. Not only does each node have multiple processors, but each typically has multiple computer cores. For many data analysis tasks, one can exploit massive amounts of parallelism at the instruction level. For example, filtering noise from sensor data may involve invoking a Fast Fourier Transform (FFT) or other spectral methods. These computations can be sped up by using general-purpose graphics processing units (GPGPUs) in each node. Depending on the rate at which a node can access data, this GPGPU-based processing may allow us to decrease the number of nodes required to meet an overall service rate.

The World Wide Web began as a loose federation of simple Web servers that each hosted scientific documents and data of interest to a relatively small community of researchers. As the number of servers grew exponentially and the global Internet matured, Web search transformed what was initially a scientific experiment into a new economic and social force. The effectiveness of search was achievable only because of the available parallelism in massive datacenters. As we enter the period in which all of science is being driven by a data explosion, cloud computing and its inherent ability to exploit parallelism at many levels has become a fundamental new enabling technology to advance human knowledge.

REFERENCES

[1] President's Information Technology Advisory Committee, "Computational Science: Ensuring America's Competitiveness," June 2005, www.nitrd.gov/pitac/reports/20050609_computational/computational.pdf.

[2] D. A. Reed, Ed., "Workshop on The Roadmap for the Revitalization of High-End Computing," June 2003, www.cra.org/reports/supercomputing.pdf.

[3] S. L. Graham, M. Snir, and C. A. Patterson, Eds., *Getting Up to Speed: The Future of Supercomputing*, Washington, D.C.: National Academies Press, 2004, www.nap.edu/openbook.php?record_id=11148.

[4] J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," OSDI'04: Sixth Symposium on Operating Systems Design and Implementation, San Francisco, CA, Dec. 2004, doi: 10.1145/1327452.1327492.

[5] Y. Yu., M. Isard, D. Fetterly, M. Budiu, Ú. Erlingsson, P. Kumar Gunda, and J. Currey, "DryadLINQ: A System for General-Purpose Distributed Data-Parallel Computing Using a High-Level Language," OSDI'08 Eighth Symposium on Operating Systems Design and Implementation.

The Impact of Workflow Tools on Data-centric Research

CAROLE GOBLE
University of Manchester

DAVID DE ROURE
University of Southampton

WE ARE IN AN ERA OF DATA-CENTRIC SCIENTIFIC RESEARCH, in which hypotheses are not only tested through directed data collection and analysis but also generated by combining and mining the pool of data already available [1-3]. The scientific data landscape we draw upon is expanding rapidly in both scale and diversity. Taking the life sciences as an example, high-throughput gene sequencing platforms are capable of generating terabytes of data in a single experiment, and data volumes are set to increase further with industrial-scale automation. From 2001 to 2009, the number of databases reported in *Nucleic Acids Research* jumped from 218 to 1,170 [4]. Not only are the datasets growing in size and number, but they are only partly coordinated and often incompatible [5], which means that discovery and integration tasks are significant challenges. At the same time, we are drawing on a broader array of data sources: modern biology draws insights from combining different types of "omic" data (proteomic, metabolomic, transcriptomic, genomic) as well as data from other disciplines such as chemistry, clinical medicine, and public health, while systems biology links multi-scale data with multi-scale mathematical models. These data encompass all types: from structured database records to published articles, raw numeric data, images, and descriptive interpretations that use controlled vocabularies.

Data generation on this scale must be matched by scalable processing methods. The preparation, management, and analysis of data are bottlenecks and also beyond the skill of many scientists. Workflows [6] provide (1) a systematic and automated means of conducting analyses across diverse datasets and applications; (2) a way of capturing this process so that results can be reproduced and the method can be reviewed, validated, repeated, and adapted; (3) a visual scripting interface so that computational scientists can create these pipelines without low-level programming concern; and (4) an integration and access platform for the growing pool of independent resource providers so that computational scientists need not specialize in each one. The workflow is thus becoming a paradigm for enabling science on a large scale by managing data preparation and analysis pipelines, as well as the preferred vehicle for computational knowledge extraction.

WORKFLOWS DEFINED

A workflow is a precise description of a scientific procedure—a multi-step process to coordinate multiple tasks, acting like a sophisticated script [7]. Each task represents the execution of a computational process, such as running a program, submitting a query to a database, submitting a job to a compute cloud or grid, or invoking a service over the Web to use a remote resource. Data output from one task is consumed by subsequent tasks according to a predefined graph topology that "orchestrates" the flow of data. Figure 1 presents an example workflow, encoded in the Taverna Workflow Workbench [8], which searches for genes by linking four publicly available data resources distributed in the U.S., Europe, and Japan: BioMart, Entrez, UniProt, and KEGG.

Workflow systems generally have three components: an execution platform, a visual design suite, and a development kit. The platform executes the workflow on behalf of applications and handles common crosscutting concerns, including (1) *invocation* of the service applications and handling the heterogeneity of data types and interfaces on multiple computing platforms; (2) *monitoring and recovery* from failures; (3) *optimization* of memory, storage, and execution, including concurrency and parallelization; (4) *data handling:* mapping, referencing, movement, streaming, and staging; (5) *logging* of processes and data provenance tracking; and (6) *security* and monitoring of access policies. Workflow systems are required to support long-running processes in volatile environments and thus must be robust and capable of fault tolerance and recovery. They also need to evolve continually to harness the growing capabilities of underlying computational and storage

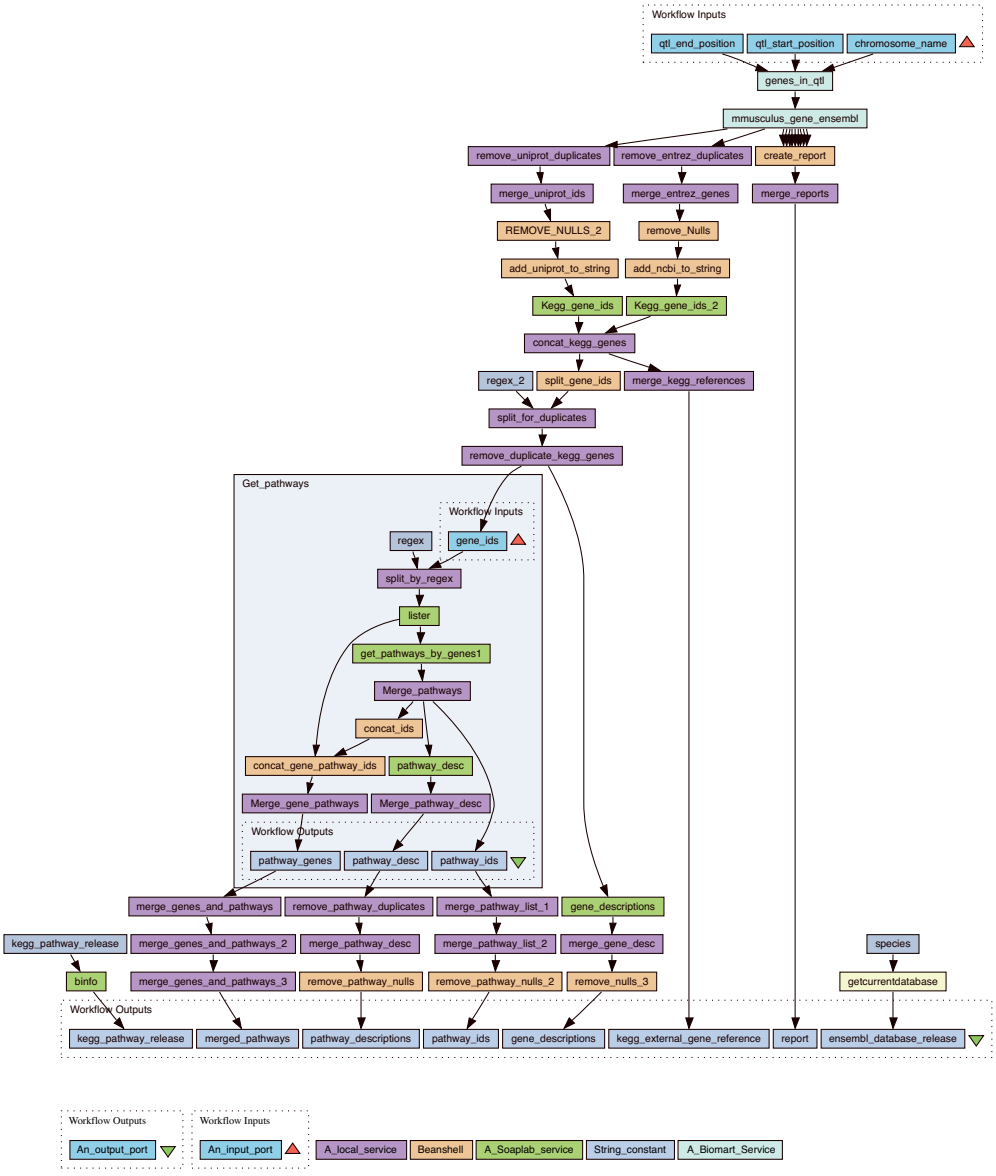


FIGURE 1.

A Taverna workflow that connects several internationally distributed datasets to identify candidate genes that could be implicated in resistance to African trypanosomiasis [11].

resources, delivering greater capacity for analysis.

The design suite provides a visual scripting application for authoring and sharing workflows and preparing the components that are to be incorporated as executable steps. The aim is to shield the author from the complexities of the underlying applications and enable the author to design and understand workflows without recourse to commissioning specialist and specific applications or hiring software engineers. This empowers scientists to build their own pipelines when they need them and how they want them. Finally, the development kit enables developers to extend the capabilities of the system and enables workflows to be embedded into applications, Web portals, or databases. This embedding is transformational: it has the potential to incorporate sophisticated knowledge seamlessly and invisibly into the tools that scientists use routinely.

Each workflow system has its own language, design suite, and software components, and the systems vary in their execution models and the kinds of components they coordinate [9]. Sedna is one of the few to use the industry-standard Business Process Execution Language (BPEL) for scientific workflows [10]. General-purpose open source workflow systems include Taverna,[1] Kepler,[2] Pegasus,[3] and Triana.[4] Other systems, such as the LONI Pipeline[5] for neuroimaging and the commercial Pipeline Pilot[6] for drug discovery, are more geared toward specific applications and are optimized to support specific component libraries. These focus on interoperating applications; other workflow systems target the provisioning of compute cycles or submission of jobs to grids. For example, Pegasus and DAGMan[7] have been used for a series of large-scale eScience experiments such as prediction models in earthquake forecasting using sensor data in the Southern California Earthquake Center (SCEC) CyberShake project.[8]

WORKFLOW USAGE

Workflows liberate scientists from the drudgery of routine data processing so they can concentrate on scientific discovery. They shoulder the burden of routine tasks, they represent the computational protocols needed to undertake data-centric

[1] www.taverna.org.uk
[2] http://kepler-project.org
[3] http://pegasus.isi.edu
[4] www.trianacode.org
[5] http://pipeline.loni.ucla.edu
[6] http://accelrys.com/products/scitegic
[7] www.cs.wisc.edu/condor/dagman
[8] http://epicenter.usc.edu/cmeportal/CyberShake.html

science, and they open up the use of processes and data resources to a much wider group of scientists and scientific application developers.

Workflows are ideal for systematically, accurately, and repeatedly running routine procedures: managing data capture from sensors or instruments; cleaning, normalizing, and validating data; securely and efficiently moving and archiving data; comparing data across repeated runs; and regularly updating data warehouses. For example, the Pan-STARRS[9] astronomical survey uses Microsoft Trident Scientific Workflow Workbench[10] workflows to load and validate telescope detections running at about 30 TB per year. Workflows have also proved useful for maintaining and updating data collections and warehouses by reacting to changes in the underlying datasets. For example, the Nijmegen Medical Centre rebuilt the tGRAP G-protein coupled receptors mutant database using a suite of text-mining Taverna workflows.

At a higher level, a workflow is an explicit, precise, and modular expression of an *in silico* or "dry lab" experimental protocol. Workflows are ideal for gathering and aggregating data from distributed datasets and data-emitting algorithms—a core activity in dataset annotation; data curation; and multi-evidential, comparative science. In Figure 1, disparate datasets are searched to find and aggregate data related to metabolic pathways implicated in resistance to African trypanosomiasis; interlinked datasets are chained together by the dataflow. In this instance, the automated and systematic processing by the workflow overcame the inadequacies of manual data triage—which leads to prematurely excluding data from analysis to cope with the quantity—and delivered new results [11].

Beyond data assembly, workflows codify data mining and knowledge discovery pipelines and parameter sweeps across predictive algorithms. For example, LEAD[11] workflows are driven by external events generated by data mining agents that monitor collections of instruments for significant patterns to trigger a storm prediction analysis; the Jet Propulsion Laboratory uses Taverna workflows for exploring a large space of multiple-parameter configurations of space instruments.

Finally, workflow systems liberate the implicit workflow embedded in an application into an explicit and reusable specification over a common software machinery and shared infrastructure. Expert informaticians use workflow systems directly as means to develop workflows for handling infrastructure; expert

[9] http://pan-starrs.ifa.hawaii.edu
[10] http://research.microsoft.com/en-us/collaboration/tools/trident.aspx
[11] http://portal.leadproject.org

scientific informaticians use them to design and explore new investigative procedures; a larger group of scientists uses precooked workflows with restricted configuration constraints launched from within applications or hidden behind Web portals.

WORKFLOW-ENABLED DATA-CENTRIC SCIENCE

Workflows offer techniques to support the new paradigm of data-centric science. They can be replayed and repeated. Results and secondary data can be computed as needed using the latest sources, providing virtual data (or on-demand) warehouses by effectively providing distributed query processing. *Smart reruns* of workflows automatically deliver new outcomes when fresh primary data and new results become available—and also when new methods become available. The workflows themselves, as first-class citizens in data-centric science, can be generated and transformed dynamically to meet the requirements at hand. In a landscape of data in considerable flux, workflows provide robustness, accountability, and full auditing. By combining workflows and their execution records with published results, we can promote systematic, unbiased, transparent, and comparable research in which outcomes carry the provenance of their derivation. This can potentially accelerate scientific discovery.

To accelerate experimental *design*, workflows can be reconfigured and repurposed as new components or templates. Creating workflows requires expertise that is hard won and often outside the skill set of the researcher. Workflows are often complex and challenging to build because they are essentially forms of programming that require some understanding of the datasets and the tools they manipulate [12]. Hence there is significant benefit in establishing shared collections of workflows that contain standard processing pipelines for immediate reuse or for repurposing in whole or in part. These aggregations of expertise and resources can help propagate techniques and best practices. Specialists can create the application steps, experts can design the workflows and set parameters, and the inexperienced can benefit by using sophisticated protocols.

The myExperiment[12] social Web site has demonstrated that by adopting content-sharing tools for repositories of workflows, we can enable social networking around workflows and provide community support for social tagging, comments, ratings and recommendations, and mixing of new workflows with those previously

[12] www.myexperiment.org

deposited [13]. This is made possible by the scale of participation in data-centric science, which can be brought to bear on challenging problems. For example, the environment of workflow execution is in such a state of flux that workflows appear to decay over time, but workflows can be kept current by a combination of expert and community curation.

Workflows enable data-centric science to be a collaborative endeavor on multiple levels. They enable scientists to collaborate over shared data and shared services, and they grant non-developers access to sophisticated code and applications without the need to install and operate them. Consequently, scientists can use the best applications, not just the ones with which they are familiar. Multidisciplinary workflows promote even broader collaboration. In this sense, a workflow system is a framework for reusing a community's tools and datasets that respects the original codes and overcomes diverse coding styles. Initiatives such as the BioCatalogue[13] registry of life science Web services and the component registries deployed at SCEC enable components to be discovered. In addition to the benefits that come from explicit sharing, there is considerable value in the information that may be gathered just through monitoring the use of data sources, services, and methods. This enables automatic monitoring of resources and recommendation of common practice and optimization.

Although the impact of workflow tools on data-centric research is potentially profound—scaling processing to match the scaling of data—many challenges exist over and above the engineering issues inherent in large-scale distributed software [14]. There are a confusing number of workflow platforms with various capabilities and purposes and little compliance with standards. Workflows are often difficult to author, using languages that are at an inappropriate level of abstraction and expecting too much knowledge of the underlying infrastructure. The reusability of a workflow is often confined to the project it was conceived in—or even to its author—and it is inherently only as strong as its components. Although workflows encourage providers to supply clean, robust, and validated data services, component failure is common. If the services or infrastructure decays, so does the workflow. Unfortunately, debugging failing workflows is a crucial but neglected topic. Contemporary workflow platforms fall short of adequately supporting rapid deployment into the user applications that consume them, and legacy application codes need to be integrated and managed.

[13] www.biocatalogue.org

CONCLUSION

Workflows affect data-centric research in four ways. First, they shift scientific practice. For example, in a data-driven hypothesis [1], data analysis yields results that are to be tested in the laboratory. Second, they have the potential to empower scientists to be the authors of their own sophisticated data processing pipelines without having to wait for software developers to produce the tools they need. Third, they offer systematic production of data that is comparable and verifiably attributable to its source. Finally, people speak of a data deluge [15], and data-centric science could be characterized as being about the primacy of data as opposed to the primacy of the academic paper or document [16], but it brings with it a method deluge: workflows illustrate *primacy of method* as another crucial paradigm in data-centric research.

REFERENCES

[1] D. B. Kell and S. G. Oliver, "Here is the evidence, now what is the hypothesis? The complementary roles of inductive and hypothesis-driven science in the post-genomic era," *BioEssays*, vol. 26, no. 1, pp. 99–105, 2004, doi: 10.1002/bies.10385.

[2] A. Halevy, P. Norvig, and F. Pereira, "The Unreasonable Effectiveness of Data," *IEEE Intell. Syst.*, vol. 24, no. 2, pp. 8–12. 2009, doi: 10.1109/MIS.2009.36.

[3] C. Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, vol. 16, no. 7, June 23, 2008, www.wired.com/science/discoveries/magazine/16-07/pb_theory.

[4] M. Y. Galperin and G. R. Cochrane, "Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009," *Nucl. Acids Res.*, vol. 37 (Database issue), pp. D1–D4, doi: 10.1093/nar/gkn942.

[5] C. Goble and R. Stevens, "The State of the Nation in Data Integration in Bioinformatics," *J. Biomed. Inform.*, vol. 41, no. 5, pp. 687–693, 2008.

[6] I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, Eds., *Workflows for e-Science: Scientific Workflows for Grids*. London: Springer, 2007.

[7] P. Romano, "Automation of in-silico data analysis processes through workflow management systems," *Brief Bioinform*, vol. 9, no. 1, pp. 57–68, Jan. 2008, doi: 10.1093/bib/bbm056.

[8] T. Oinn, M. Greenwood, M. Addis, N. Alpdemir, J. Ferris, K. Glover, C. Goble, A. Goderis, D. Hull, D. Marvin, P. Li, P. Lord, M. Pocock, M. Senger, R. Stevens, A. Wipat, and C. Wroe, "Taverna: lessons in creating a workflow environment for the life sciences," *Concurrency and Computation: Practice and Experience*, vol. 18, no. 10, pp. 1067–1100, 2006, doi: 10.1002/cpe.v18:10.

[9] E. Deelman, D. Gannon, M. Shields, and I. Taylor, "Workflows and e-Science: An overview of workflow system features and capabilities," *Future Gen. Comput. Syst.*, vol. 25, no. 5, pp. 528–540, May 2009, doi: 10.1016/j.future.2008.06.012.

[10] B. Wassermann, W. Emmerich, B. Butchart, N. Cameron, L. Chen, and J. Patel, "Sedna: a BPEL-based environment for visual scientific workflow modelling," in I. J. Taylor, E. Deelman, D. B. Gannon, and M. Shields, Eds., *Workflows for e-Science: Scientific Workflows for Grids*. London: Springer, 2007, pp. 428–449, doi: 10.1.1.103.7892.

[11] P. Fisher, C. Hedeler, K. Wolstencroft, H. Hulme, H. Noyes, S. Kemp, R. Stevens, and A. Brass,

"A Systematic Strategy for Large-Scale Analysis of Genotype-Phenotype Correlations: Identification of candidate genes involved in African Trypanosomiasis," *Nucleic Acids Res.*, vol. 35, no. 16, pp. 5625–5633, 2007, doi: 10.1093/nar/gkm623.

[12] A. Goderis, U. Sattler, P. Lord, and C. Goble, "Seven Bottlenecks to Workflow Reuse and Repurposing in The Semantic Web," *ISWC 2005*, pp. 323–337, doi: 10.1007/11574620_25.

[13] D. De Roure, C. Goble, and R. Stevens, "The Design and Realisation of the myExperiment Virtual Research Environment for Social Sharing of Workflows," *Future Gen. Comput. Syst.*, vol. 25, pp. 561–567, 2009, doi: 10.1016/j.future.2008.06.010.

[14] Y. Gil, E. Deelman, M. Ellisman, T. Fahringer, G. Fox, D. Gannon, C. Goble, M. Livny, L. Moreau, and J. Myers, "Examining the Challenges of Scientific Workflows," *Computer*, vol. 40, pp. 24–32, 2007, doi: 10.1109/MC.2007.421.

[15] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science*, vol. 323, no. 5919, pp. 1297–1298, Mar. 6, 2009, doi: 10.1126/science.1170411.

[16] G. Erbach, "Data-centric view in e-Science information systems," *Data Sci. J.*, vol. 5, pp. 219–222, 2006, doi: 10.2481/dsj.5.219.

Semantic eScience: Encoding Meaning in Next-Generation Digitally Enhanced Science

PETER FOX
JAMES HENDLER
Rensselaer Polytechnic Institute

Science is becoming increasingly dependent on data, yet traditional data technologies were not designed for the scale and heterogeneity of data in the modern world. Projects such as the Large Hadron Collider (LHC) and the Australian Square Kilometre Array Pathfinder (ASKAP) will generate petabytes of data that must be analyzed by hundreds of scientists working in multiple countries and speaking many different languages. The digital or electronic facilitation of science, or eScience [1], is now essential and becoming widespread.

Clearly, data-intensive science, one component of eScience, must move beyond data warehouses and closed systems, striving instead to allow access to data to those outside the main project teams, allow for greater integration of sources, and provide interfaces to those who are expert scientists but not experts in data administration and computation. As eScience flourishes and the barriers to free and open access to data are being lowered, other, more challenging, questions are emerging, such as, "How do I use this data that I did not generate?" or "How do I use this data type, which I have never seen, with the data I use every day?" or "What should I do if I really need data from another discipline but I cannot understand its terms?" This list of questions is large and growing as data and information product use increases and as more of science comes to rely on specialized devices.

An important insight into dealing with heterogeneous data is that if you know what the data "means," it will be easier to use. As the volume, complexity, and heterogeneity of data resources grow, scientists increasingly need new capabilities that rely on new "semantic" approaches (e.g., in the form of ontologies—machine encodings of terms, concepts, and relations among them). Semantic technologies are gaining momentum in eScience areas such as solar-terrestrial physics (see Figure 1), ecology,[1] ocean and marine sciences,[2] healthcare, and life sciences,[3] to name but a few. The developers of eScience infrastructures are increasingly in need of semantic-based methodologies, tools, and middleware. They can in turn facilitate scientific knowledge modeling, logic-based hypothesis checking, semantic data integration, application composition, and integrated knowledge discovery and data analysis for different scientific domains and systems noted above, for use by scientists, students, and, increasingly, non-experts.

The influence of the artificial intelligence community and the increasing amount of data available on the Web (which has led many scientists to use the Web as their primary "computer") have led semantic Web researchers to focus both on formal aspects of semantic representation languages and on general-purpose semantic application development. Languages are being standardized, and communities are in turn using those languages to build and use ontologies—specifications of concepts and terms and the relations between them (in the formal, machine-readable sense). All of the capabilities currently needed by eScience—including data integration, fusion, and mining; workflow development, orchestration, and execution; capture of provenance, lineage, and data quality; validation, verification, and trust of data authenticity; and fitness for purpose—need semantic representation and mediation if eScience is to become fully data-intensive.

The need for more semantics in eScience also arises in part from the increasingly distributed and interdisciplinary challenges of modern research. For example, the availability of high spatial-resolution remote sensing data (such as imagery) from satellites for ecosystem science is simultaneously changing the nature of research in other scientific fields, such as environmental science. Yet ground-truthing with *in situ* data creates an immediate data-integration challenge. Questions that arise for researchers who use such data include, "How can 'point' data be reconciled with various satellite data—e.g., swath or gridded—products?" "How is the spatial

[1] E.g., the Science Environment for Ecological Knowledge (SEEK) and [2].
[2] E.g., the Marine Metadata Interoperability (MMI) project.
[3] E.g., the Semantic Web Health Care and Life Sciences (HCLS) Interest Group and [3].

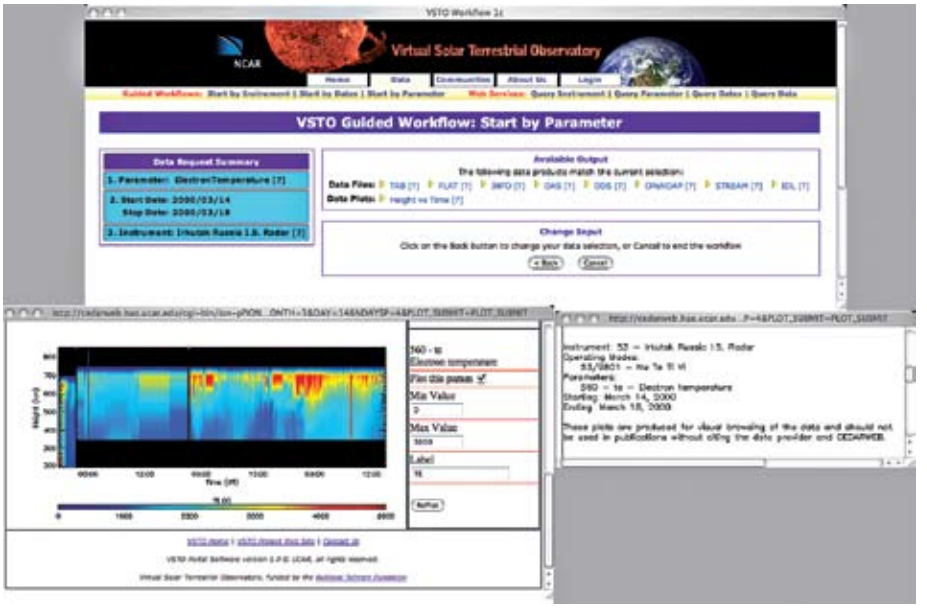


FIGURE 1.
The Virtual Solar-Terrestrial Observatory (VSTO) provides data integration between physical parameters measured by different instruments. VSTO also mediates independent coordinate information to select appropriate plotting types using a semantic eScience approach without the user having to know the underlying representations and structure of the data [4, 5].

registration performed?" "Do these data represent the 'same' thing, at the same vertical (as well as geographic) position or at the same time, and does that matter?" Another scientist, such as a biologist, might need to access the same data from a very different perspective, to ask questions such as, "I found this particular species in an unexpected location. What are the geophysical parameters—temperature, humidity, and so on—for this area, and how has it changed over the last weeks, months, years?" Answers to such questions reside in both the metadata and the data itself. Perhaps more important is the fact that data and information products are increasingly being made available via Web services, so the semantic binding (i.e., the meaning) we seek must shift from being at the data level to being at the Internet/Web service level.

Semantics adds not only well-defined and machine-encoded definitions of vo-

cabularies, concepts, and terms, but it also explains the interrelationships among them (and especially, on the Web, among different vocabularies residing in different documents or repositories) in declarative (stated) and conditional (e.g., rule-based or logic) forms. One of the present challenges around semantic eScience is balancing expressivity (of the semantic representation) with the complexity of defining terms used by scientific experts and implementing the resulting systems. This balance is application dependent, which means there is no one-approach-fits-all solution. In turn, this implies that a peer relationship is required between physical scientists and computer scientists, and between software engineers and data managers and data providers.

The last few years have seen significant development in Web-based (i.e., XML) markup languages, including stabilization and standardization. Retrospective data and their accompanying catalogs are now provided as Web services, and real-time and near-real-time data are becoming standardized as sensor Web services are emerging. This means that diverse datasets are now widely available. Clearinghouses for such service registries, including the Earth Observing System Clearinghouse (ECHO) and the Global Earth Observation System of Systems (GEOSS) for Earth science, are becoming populated, and these complement comprehensive inventory catalogs such as NASA's Global Change Master Directory (GCMD). However, these registries remain largely limited to syntax-only representations of the services and underlying data. Intensive human effort—to match inputs, outputs, and preconditions as well as the meaning of methods for the services—is required to utilize them.

Project and community work to develop data models to improve lower-level interoperability is also increasing. These models expose domain vocabularies, which is helpful for immediate domains of interest but not necessarily for crosscutting areas such as Earth science data records and collections. As noted in reports from the international level to the agency level, data from new missions, together with data from existing agency sources, are increasingly being used synergistically with other observing and modeling sources. As these data sources are made available as services, the need for interoperability among differing vocabularies, services, and method representations remains, and the limitations of syntax-only (or lightweight semantics, such as coverage) become clear. Further, as demand for information products (representations of the data beyond pure science use) increases, the need for non-specialist access to information services based on science data is rapidly increasing. This need is not being met in most application areas.

Those involved in extant efforts (noted earlier, such as solar-terrestrial physics,

ecology, ocean and marine sciences, healthcare, and life sciences) have made the case for interoperability that moves away from reliance on agreements at the data-element, or syntactic, level toward a higher scientific, or semantic, level. Results from such research projects have demonstrated these types of data integration capabilities in interdisciplinary and cross-instrument measurement use. Now that syntax-only interoperability is no longer state-of-the-art, the next logical step is to use the semantics to begin to enable a similar level of semantic support at the data-as-a-service level.

Despite this increasing awareness of the importance of semantics to data-intensive eScience, participation from the scientific community to develop the particular requirements from specific science areas has been inadequate. Scientific researchers are growing ever more dependent on the Web for their data needs, but to date they have not yet created a coherent agenda for exploring the emerging trends being enabled by semantic technologies and for interacting with Semantic Web researchers. To help create such an agenda, we need to develop a multi-disciplinary field of *semantic eScience* that fosters the growth and development of data-intensive scientific applications based on semantic methodologies and technologies, as well as related knowledge-based approaches. To this end, we issue a four-point call to action:

- Researchers in science must work with colleagues in computer science and informatics to develop field-specific requirements and to implement and evaluate the languages, tools, and applications being developed for semantic eScience.
- Scientific and professional societies must provide the settings in which the needed rich interplay between science requirements and informatics capabilities can be realized, and they must acknowledge the importance of this work in career advancement via citation-like metrics.
- Funding agencies must increasingly target the building of communities of practice, with emphasis on the types of interdisciplinary teams of researchers and practitioners that are needed to advance and sustain semantic eScience efforts.
- All parties—scientists, societies, and funders—must play a role in creating governance around controlled vocabularies, taxonomies, and ontologies that can be used in scientific applications to ensure the currency and evolution of knowledge encoded in semantics.

Although early efforts are under way in all four areas, much more must be done. The very nature of dealing with the increasing complexity of modern science demands it.

REFERENCES

[1] T. Hey and A. E. Trefethen, "Cyberinfrastructure for e-Science," *Science*, vol. 308, no. 5723, May 2005, pp. 817–821, doi: 10.1126/science.1110410.

[2] J. Madin, S. Bowers, M. Schildhauer, S. Krivov, D. Pennington, and F. Villa, "An Ontology for Describing and Synthesizing Ecological Observation Data," *Ecol. Inf.*, vol. 2, no. 3, pp. 279–296, 2007, doi: 10.1016/j.ecoinf.2007.05.004.

[3] E. Neumann, "A Life Science Semantic Web: Are We There Yet?" *Sci. STKE*, p. 22, 2005, doi: 10.1126/stke.2832005pe22.

[4] P. Fox, D. McGuinness, L. Cinquini, P. West, J. Garcia, and J. Benedict, "Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience," *Comput. Geosci.*, vol. 35, no. 4, pp. 724–738, 2009, doi: 10.1.1.141.1827.

[5] D. McGuinness, P. Fox, L. Cinquini, P. West, J. Garcia, J. L. Benedict, and D. Middleton, "The Virtual Solar-Terrestrial Observatory: A Deployed Semantic Web Application Case Study for Scientific Research," *AI Mag.*, vol. 29, no. 1 , pp. 65–76, 2007, doi: 10.1145/1317353.1317355.

Visualization for Data-Intensive Science

CHARLES HANSEN
CHRIS R. JOHNSON
VALERIO PASCUCCI
CLAUDIO T. SILVA
University of Utah

Since the advent of computing, the world has experienced an information "big bang": an explosion of data. The amount of information being created is increasing at an exponential rate. Since 2003, digital information has accounted for 90 percent of all information produced [1], vastly exceeding the amount of information on paper and on film. One of the greatest scientific and engineering challenges of the 21st century will be to understand and make effective use of this growing body of information. Visual data analysis, facilitated by interactive interfaces, enables the detection and validation of expected results while also enabling unexpected discoveries in science. It allows for the validation of new theoretical models, provides comparison between models and datasets, enables quantitative and qualitative querying, improves interpretation of data, and facilitates decision making. Scientists can use visual data analysis systems to explore "what if" scenarios, define hypotheses, and examine data using multiple perspectives and assumptions. They can identify connections among large numbers of attributes and quantitatively assess the reliability of hypotheses. In essence, visual data analysis is an integral part of scientific discovery and is far from a solved problem. Many avenues for future research remain open. In this article, we describe visual data analysis topics that will receive attention in the next decade [2, 3].

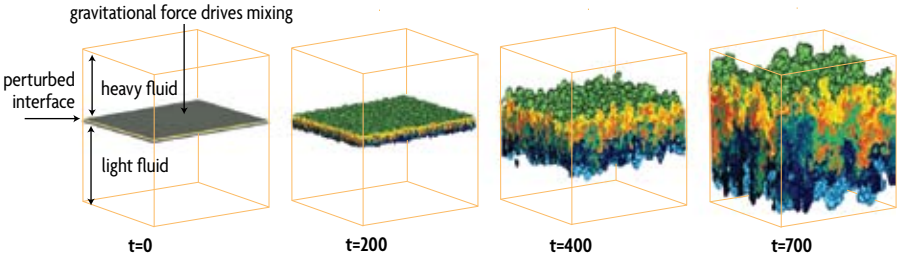


FIGURE 1.
Interactive visualization of four timesteps of the 1152^3 simulation of a Rayleigh-Taylor instability. Gravity drives the mixing of a heavy fluid on top of a lighter one. Two envelope surfaces capture the mixing region.

VISUS: PROGRESSIVE STREAMING FOR SCALABLE DATA EXPLORATION

In recent years, computational scientists with access to the world's largest supercomputers have successfully simulated a number of natural and man-made phenomena with unprecedented levels of detail. Such simulations routinely produce massive amounts of data. For example, hydrodynamic instability simulations performed at Lawrence Livermore National Laboratory (LLNL) in early 2002 produced several tens of terabytes of data, as shown in Figure 1. This data must be visualized and analyzed to verify and validate the underlying model, understand the phenomenon in detail, and develop new insights into its fundamental physics. Therefore, both visualization and data analysis algorithms require new, advanced designs that enable high performance when dealing with large amounts of data.

Data-streaming techniques and out-of-core computing specifically address the issues of algorithm redesign and data layout restructuring, which are necessary to enable scalable processing of massive amounts of data. For example, space-filling curves have been used to develop a static indexing scheme called ViSUS,[1] which produces a data layout that enables the hierarchical traversal of *n*-dimensional regular grids. Three features make this approach particularly attractive: (1) the order of the data is independent of the parameters of the physical hardware (a cache-oblivious approach), (2) conversion from Z-order used in classical database approaches is achieved using a simple sequence of bit-string manipulations, and (3) it does not introduce any data replication. This approach has

[1] www.pascucci.org/visus

FIGURE 2.
Scalability of the ViSUS infrastructure, which is used for visualization in a variety of applications (such as medical imaging, subsurface modeling, climate modeling, microscopy, satellite imaging, digital photography, and large-scale scientific simulations) and with a wide range of devices (from the iPhone to the powerwall).

been used for direct streaming and real-time monitoring of large-scale simulations during execution [4].

Figure 2 shows the ViSUS streaming infrastructure streaming LLNL simulation codes and visualizing them in real time on the Blue Gene/L installation at the Supercomputing 2004 exhibit (where Blue Gene/L was introduced as the new fastest supercomputer in the world). The extreme scalability of this approach allows the use of the same code base for a large set of applications while exploiting a wide range of devices, from large powerwall displays to workstations, laptop computers, and handheld devices such as the iPhone.

Generalization of this class of techniques to the case of unstructured meshes remains a major problem. More generally, the fast evolution and growing diversity of hardware pose a major challenge in the design of software infrastructures that are intrinsically scalable and adaptable to a variety of computing resources and running conditions. This poses theoretical and practical questions that future researchers in visualization and analysis for data-intensive applications will need to address.

VISTRAILS: PROVENANCE AND DATA EXPLORATION

Data exploration is an inherently creative process that requires the researcher to locate relevant data, visualize the data and discover relationships, collaborate with

peers while exploring solutions, and disseminate results. Given the volume of data and complexity of analyses that are common in scientific exploration, new tools are needed and existing tools should be extended to better support creativity.

The ability to systematically capture provenance is a key requirement for these tools. The provenance (also referred to as the audit trail, lineage, or pedigree) of a data product contains information about the process and data used to derive the data product. The importance of keeping provenance for data products is well recognized in the scientific community [5, 6]. It provides important documentation that is key to preserving the data, determining its quality and authorship, and reproducing and validating the results. The availability of provenance also supports reflective reasoning, allowing users to store temporary results, make inferences from stored knowledge, and follow chains of reasoning backward and forward.

VisTrails[2] is an open source system that we designed to support exploratory computational tasks such as visualization, data mining, and integration. VisTrails provides a comprehensive provenance management infrastructure and can be easily combined with existing tools and libraries. A new concept we introduced with VisTrails is the notion of *provenance of workflow evolution* [7]. In contrast to previous workflow and visualization systems, which maintain provenance only for derived data products, VisTrails treats the workflows (or pipelines) as first-class data items and keeps their provenance. VisTrails is an extensible system. Like workflow systems, it allows pipelines to be created that combine multiple libraries. In addition, the VisTrails provenance infrastructure can be integrated with interactive tools, which cannot be easily wrapped in a workflow system [8].

Figure 3 shows an example of an exploratory visualization using VisTrails. In the center, the visual trail, or *vistrail*, captures all modifications that users apply to the visualizations. Each node in the vistrail tree corresponds to a pipeline, and the edges between two nodes correspond to changes applied to transform the parent pipeline into the child (e.g., through the addition of a module or a change to a parameter value). The tree-based representation allows a scientist to return to a previous version in an intuitive way, undo bad changes, compare workflows, and be reminded of the actions that led to a particular result.

Ad hoc approaches to data exploration, which are widely used in the scientific community, have serious limitations. In particular, scientists and engineers need

[2] http://vistrails.sci.utah.edu

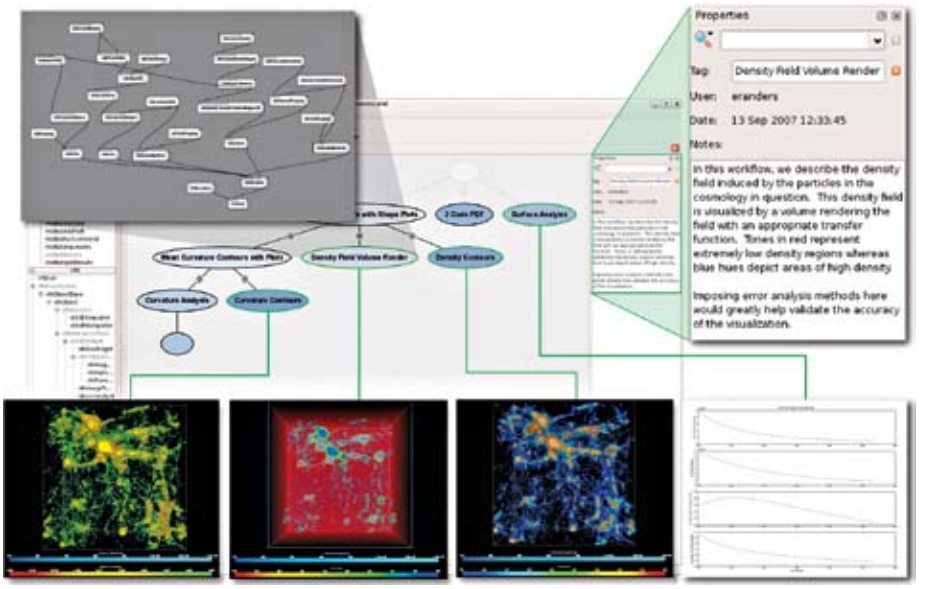


FIGURE 3.
An example of an exploratory visualization for studying celestial structures derived from cosmological simulations using VisTrails. Complete provenance of the exploration process is displayed as a "vistrail." Detailed metadata are also stored, including free-text notes made by the scientist, the date and time the workflow was created or modified, optional descriptive tags, and the name of the person who created it.

to expend substantial effort managing data (e.g., scripts that encode computational tasks, raw data, data products, images, and notes) and need to record provenance so that basic questions can be answered, such as: Who created the data product and when? When was it modified, and by whom? What process was used to create it? Were two data products derived from the same raw data? This process is not only time consuming but error prone. The absence of provenance makes it hard (and sometimes impossible) to reproduce and share results, solve problems collaboratively, validate results with different input data, understand the process used to solve a particular problem, and reuse the knowledge involved in the data analysis process. It also greatly limits the longevity of the data product. Without precise and sufficient information about how it was generated, its value is greatly diminished. Visualization systems aimed at the scientific domain need to provide a flexible

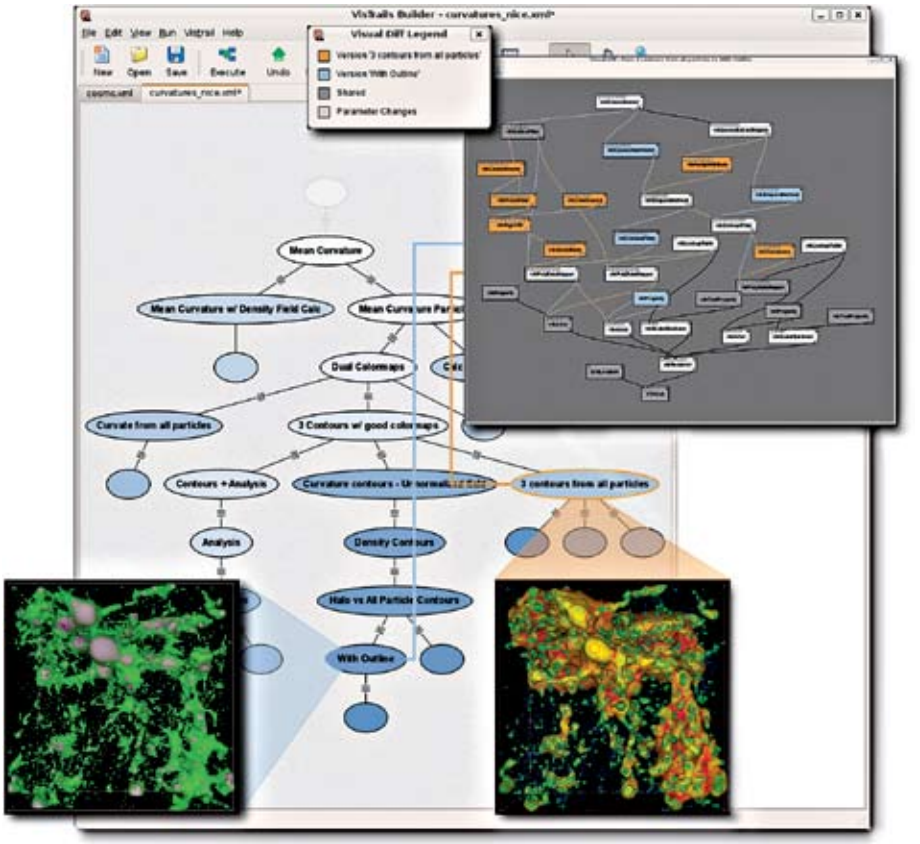


FIGURE 4.
Representing provenance as a series of actions that modify a pipeline makes visualizing the differences between two workflows possible. The difference between two workflows is represented in a meaningful way, as an aggregation of the two. This is both informative and intuitive, reducing the time it takes to understand how two workflows are functionally different.

framework that not only enables scientists to perform complex analyses over large datasets but also captures detailed provenance of the analysis process.

Figure 4 shows ParaView[3] (a data analysis and visualization tool for extreme-

[3] www.paraview.org

ly large datasets) and the VisTrails Provenance Explorer transparently capturing a complete exploration process. The provenance capture mechanism was implemented by inserting monitoring code in ParaView's undo/redo mechanism, which captures changes to the underlying pipeline specification. Essentially, the action on top of the undo stack is added to the vistrail in the appropriate place, and undo is reinterpreted to mean "move up the version tree." Note that the change-based representation is both simple and compact—it uses substantially less space than the alternative approach of storing multiple instances, or versions, of the state.

FLOW VISUALIZATION TECHNIQUES

A precise qualitative and quantitative assessment of three-dimensional transient flow phenomena is required in a broad range of scientific, engineering, and medical applications. Fortunately, in many cases the analysis of a 3-D vector field can be reduced to the investigation of the two-dimensional structures produced by its interaction with the boundary of the object under consideration. Typical examples of such analysis for fluid flows include airfoils and reactors in aeronautics, engine walls and exhaust pipes in the automotive industry, and rotor blades in turbomachinery.

Other applications in biomedicine focus on the interplay between bioelectric fields and the surface of an organ. In each case, numerical simulations of increasing size and sophistication are becoming instrumental in helping scientists and engineers reach a deeper understanding of the flow properties that are relevant to their task. The scientific visualization community has concentrated a significant part of its research efforts on the design of visualization methods that convey local and global structures that occur at various spatial and temporal scales in transient flow simulations. In particular, emphasis has been placed on the interactivity of the corresponding visual analysis, which has been identified as a critical aspect of the effectiveness of the proposed algorithms.

A recent trend in flow visualization research is to use GPUs to compute image space methods to tackle the computational complexity of visualization techniques that support flows defined over curved surfaces. The key feature of this approach is the ability to efficiently produce a dense texture representation of the flow without explicitly computing a surface parameterization. This is achieved by projecting onto the image plane the flow corresponding to the visible part of the surface, allowing subsequent texture generation in the image space through backward integration and iterative blending. Although the use of partial surface parameterization obtained by projection results in an impressive performance gain, texture patterns

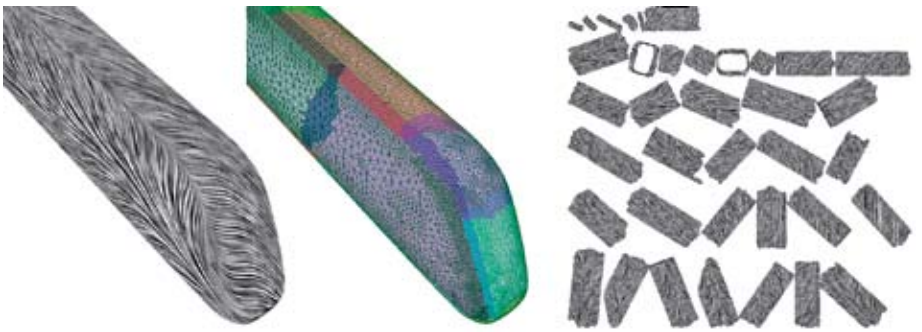


FIGURE 5.
Simulation of a high-speed ICE train. Left: The GPUFLIC result. Middle: Patch configurations. Right: Charts in texture space.

stretching beyond the visible part of the self-occluded surface become incoherent due to the lack of full surface parameterization.

To address this problem, we have introduced a novel scheme that fully supports the creation of high-quality texture-based visualizations of flows defined over arbitrary curved surfaces [9]. Called Flow Charts, our scheme addresses the issue mentioned previously by segmenting the surface into overlapping patches, which are then individually parameterized into charts and packed in the texture domain. The overlapped region provides each local chart with a smooth representation of its direct vicinity in the flow domain as well as with the inter-chart adjacency information, both of which are required for accurate and non-disrupted particle advection. The vector field and the patch adjacency relation are naturally represented as textures, enabling efficient GPU implementation of state-of-the-art flow texture synthesis algorithms such as GPUFLIC and UFAC.

Figure 5 shows the result of a simulation of a high-speed German Intercity-Express (ICE) train traveling at a velocity of about 250 km/h with wind blowing from the side at an incidence angle of 30 degrees. The wind causes vortices to form on the lee side of the train, creating a drop in pressure that adversely affects the train's ability to stay on the track. These flow structures induce separation and attachment flow patterns on the train surface. They can be clearly seen in the proposed images close to the salient edges of the geometry.

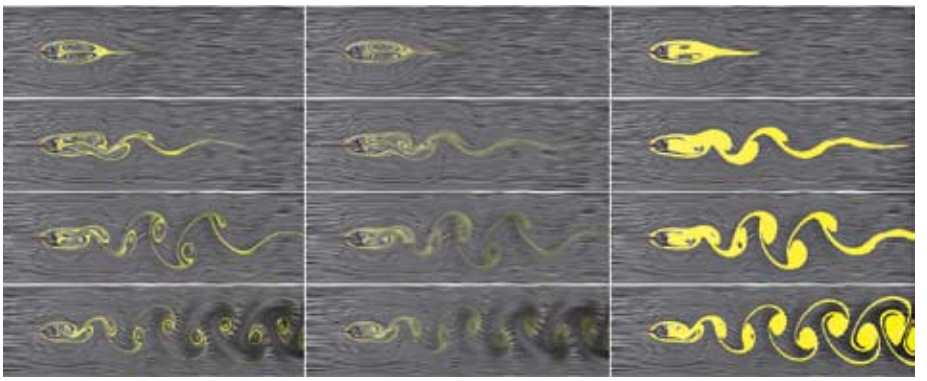


FIGURE 6.
Visualization of the Karman dataset using dye advection. Left column: Physically based dye advection. Middle column: Texture advection method. Right column: Level-set method. The time sequence is from top to bottom.

The effectiveness of a physically based formulation can be seen with the Karman dataset (Figure 6), a numerical simulation of the classical Von Kármán vortex street phenomenon, in which a repeating pattern of swirling vortices is caused by the separation of flow passing over a circular-shaped obstacle. The visualization of dye advection is overlaid on dense texture visualization that shows instantaneous flow structures generated by GPUFLIC. The patterns generated by the texture-advection method are hazy due to numerical diffusion and loss of mass. In a level-set method, intricate structures are lost because of the binary dye/background threshold. Thanks to the physically based formulation [10], the visualization is capable of accurately conveying detailed structures not shown using the traditional texture-advection method.

FUTURE DATA-INTENSIVE VISUALIZATION CHALLENGES

Fundamental advances in visualization techniques and systems must be made to extract meaning from large and complex datasets derived from experiments and from upcoming petascale and exascale simulation systems. Effective data analysis and visualization tools in support of predictive simulations and scientific knowledge discovery must be based on strong algorithmic and mathematical foundations

and must allow scientists to reliably characterize salient features in their data. New mathematical methods in areas such as topology, high-order tensor analysis, and statistics will constitute the core of feature extraction and uncertainty modeling using formal definition of complex shapes, patterns, and space-time distributions. Topological methods are becoming increasingly important in the development of advanced data analysis because of their expressive power in describing complex shapes at multiple scales. The recent introduction of robust combinatorial techniques for topological analysis has enabled the use of topology—not only for presentation of known phenomena but for the detection and quantification of new features of fundamental scientific interest.

Our current data-analysis capabilities lag far behind our ability to produce simulation data or record observational data. New visual data analysis techniques will need to dynamically consider high-dimensional probability distributions of quantities of interest. This will require new contributions from mathematics, probability, and statistics. The scaling of simulations to ever-finer granularity and timesteps brings new challenges in visualizing the data that is generated. It will be crucial to develop smart, semi-automated visualization algorithms and methodologies to help filter the data or present "summary visualizations" to enable scientists to begin analyzing the immense datasets using a more top-down methodological path. The ability to fully quantify uncertainty in high-performance computational simulations will provide new capabilities for verification and validation of simulation codes. Hence, uncertainty representation and quantification, uncertainty propagation, and uncertainty visualization techniques need to be developed to provide scientists with credible and verifiable visualizations.

New approaches to visual data analysis and knowledge discovery are needed to enable researchers to gain insight into this emerging form of scientific data. Such approaches must take into account the multi-model nature of the data; provide the means for scientists to easily transition views from global to local model data; allow blending of traditional scientific visualization and information visualization; perform hypothesis testing, verification, and validation; and address the challenges posed by the use of vastly different grid types and by the various elements of the multi-model code. Tools that leverage semantic information and hide details of dataset formats will be critical to enabling visualization and analysis experts to concentrate on the design of these approaches rather than becoming mired in the trivialities of particular data representations [11].

ACKNOWLEDGMENTS

Publication is based, in part, on work supported by DOE: VACET, DOE SDM, DOE C-SAFE Alliance Center, the National Science Foundation (grants IIS-0746500, CNS-0751152, IIS-0713637, OCE-0424602, IIS-0534628, CNS-0514485, IIS-0513692, CNS-0524096, CCF-0401498, OISE-0405402, CNS-0615194, CNS-0551724, CCF-0541113, IIS-0513212, and CCF-0528201), IBM Faculty Awards (2005, 2006, and 2007), NIH NCRR Grant No. 5P41RR012553-10 and Award Number KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors would like to thank Juliana Freire and the VisTrails team for help with the third section of this article.

REFERENCES

[1] C. R. Johnson, R. Moorhead, T. Munzner, H. Pfister, P. Rheingans, and T. S. Yoo, Eds., *NIH-NSF Visualization Research Challenges Report*, IEEE Press, ISBN 0-7695-2733-7, 2006, http://vgtc.org/wpmu/techcom/national-initiatives/nihnsf-visualization-research-challenges-report-january-2006, doi: 10.1109/MCG.2006.44.

[2] NSF Blue Ribbon Panel Report on Simulation-Based Engineering Science (J. T. Oden, T. Belytschko, J. Fish, T. Hughes, C. R. Johnson, D. Keyes, A. Laub, L. Petzold, D. Srolovitz, and S. Yip), "Simulation-Based Engineering Science," 2006, www.nd.edu/~dddas/References/SBES_Final_Report.pdf.

[3] NIH-NSF Visualization Research Challenges, http://erie.nlm.nih.gov/evc/meetings/vrc2004.

[4] V. Pascucci, D. E. Laney, R. J. Frank, F. Gygi, G. Scorzelli, L. Linsen, and B. Hamann, "Real-time monitoring of large scientific simulations," *SAC*, pp. 194–198, ACM, 2003, doi: 10.1.1.66.9717.

[5] S. B. Davidson and J. Freire, "Provenance and scientific workflows: challenges and opportunities," *Proc. ACM SIGMOD*, pp. 1345–1350, 2008, doi: 10.1.1.140.3264.

[6] J. Freire, D. Koop, E. Santos, and C. Silva, "Provenance for computational tasks: A survey," *Comput. Sci. Eng*, vol. 10, no. 3, pp. 11–21, 2008, doi: 10.1109/MCSE.2008.79.

[7] J. Freire, C. T. Silva, S. P. Callahan, E. Santos, C. E. Scheidegger, and H. T. Vo, "Managing rapidly-evolving scientific workflows," International Provenance and Annotation Workshop (IPAW), LNCS 4145, pp. 10–18, 2006, doi: 10.1.1.117.5530.

[8] C. Silva, J. Freire, and S. P. Callahan, "Provenance for visualizations: Reproducibility and beyond," *IEEE Comput. Sci. Eng.*, 2007, doi: 10.1109/MCSE.2007.106.

[9] G.-S. Li, X. Tricoche, D. Weiskopf, and C. Hansen, "Flow charts: Visualization of vector fields on arbitrary surfaces," *IEEE Trans. Visual. Comput. Graphics*, vol. 14, no. 5, pp. 1067–1080, 2008, doi: 10.1109/TVCG.2008.58.

[10] G.-S. Li, C. Hansen, and X. Tricoche, "Physically-based dye advection for flow visualization. *Comp. Graphics Forum J.*, vol. 27, no. 3, pp. 727–735, 2008, doi: 10.1111/j.1467-8659.2008.01201.x.

[11] "Visualization and Knowledge Discovery: Report from the DOE/ASCR Workshop on Visual Analysis and Data Exploration at Extreme Scale," C. R. Johnson, R. Ross, S. Ahern, J. Ahrens, W. Bethel, K. L. Ma, M. Papka, J. van Rosendale, H. W. Shen, and J. Thomas, www.sci.utah.edu/vaw2007/DOE-Visualization-Report-2007.pdf, 2007.

A Platform for All That We Know: Creating a Knowledge-Driven Research Infrastructure

SAVAS PARASTATIDIS
Microsoft

Computer systems have become a vital part of the modern research environment, supporting all aspects of the research lifecycle [1]. The community uses the terms "eScience" and "eResearch" to highlight the important role of computer technology in the ways we undertake research, collaborate, share data and documents, submit funding applications, use devices to automatically and accurately collect data from experiments, deploy new generations of microscopes and telescopes to increase the quality of the acquired imagery, and archive everything along the way for provenance and long-term preservation [2, 3].

However, the same technological advances in data capture, generation, and sharing and the automation enabled by computers have resulted in an unprecedented explosion in data—a situation that applies not only to research but to every aspect of our digital lives. This data deluge, especially in the scientific domain, has brought new research infrastructure challenges, as highlighted by Jim Gray and Alex Szalay [4]. The processing, data transfer, and storage demands are far greater today than just a few years ago. It is no surprise that we are talking about the emergence of a new research methodology—the "fourth paradigm"—in science.

THE FOURTH PARADIGM

Through the use of technology and automation, we are trying to keep up with the challenges of the data deluge. The emergence of the Web as an application, data sharing, and collaboration platform has broken many barriers in the way research is undertaken and disseminated. The emerging cloud computing infrastructures (e.g., Amazon's[1]) and the new generation of data-intensive computing platforms (e.g., DISC,[2] Google's MapReduce,[3] Hadoop,[4] and Dryad[5]) are geared toward managing and processing large amounts of data. Amazon is even offering a "sneakernet"[6]-like service[7] to address the problem of transferring large amounts of data into its cloud. Companies such as Google, Yahoo!, and Microsoft are demonstrating that it is possible to aggregate huge amounts of data from around the Web and store, manage, and index it and then build engaging user experiences around it.

The primary focus of the current technologies addresses only the first part of the data-information-knowledge-wisdom spectrum.[8] Computers have become efficient at storing, managing, indexing, and computing (research) data. They are even able to represent and process some of the information hidden behind the symbols used to encode that data. Nevertheless, we are still a long way from having computer systems that can automatically discover, acquire, organize, analyze, correlate, interpret, infer, and reason over information that's on the Internet, that's hidden on researchers' hard drives, or that exists only in our brains. We do not yet have an infrastructure capable of managing and processing knowledge on a global scale, one that can act as the foundation for a generation of knowledge-driven services and applications.

So, if the fourth paradigm is about data and information, it is not unreasonable to foresee a future, not far away, where we begin thinking about the challenges of managing knowledge and machine-based understanding on a very large scale. We researchers will probably be the first to face this challenge.

[1] http://aws.amazon.com
[2] www.pdl.cmu.edu/DISC
[3] http://labs.google.com/papers/mapreduce.html
[4] http://hadoop.apache.org
[5] http://research.microsoft.com/en-us/projects/dryad
[6] http://en.wikipedia.org/wiki/Sneakernet
[7] http://aws.amazon.com/importexport
[8] http://en.wikipedia.org/wiki/DIKW

KNOWLEDGE-ORIENTED RESEARCH INFRASTRUCTURES

The work by the Semantic Web[9] community has resulted in a number of technologies to help with data modeling, information representation, and the interexchange of semantics, always within the context of a particular application domain. Given the formal foundations of some of these technologies (e.g., the Web Ontology Language, or OWL), it has been possible to introduce reasoning capabilities, at least for some specific bounded domains (e.g., BioMoby[10]).

Moving forward, the work of the Semantic Web community will continue to play a significant role in the interoperable exchange of information and knowledge. More importantly, as representation technologies such as RDF (Resource Description Framework), OWL, and microformats become widely accepted, the focus will transition to the computational aspects of semantic understanding and knowledge. The challenge we will face is the automation of the aggregation and combination of huge amounts of semantically rich information and, very crucially, the processes by which that information is generated and analyzed. Today, we must start thinking about the technologies we'll need in order to semantically describe, analyze, and combine the information and the algorithms used to produce it or consume it, and to do so on a global scale. If today's cloud computing services focus on offering a scalable platform for computing, tomorrow's services will be built around the management of knowledge and reasoning over it.

We are already seeing some attempts to infer knowledge based on the world's information. Services such as OpenCyc,[11] Freebase,[12] Powerset,[13] True Knowledge,[14] and Wolfram|Alpha[15] demonstrate how facts can be recorded in such a way that they can be combined and made available as answers to a user's questions. Wolfram|Alpha, in particular, has made use of domain experts to encode the computational aspects of processing the data and information that they have aggregated from around the Web and annotated. It demonstrates how a consumer-oriented service can be built on top of a computational infrastructure in combination with natural language processing. It is likely that many similar services will emerge in the near future, initially targeting specialized technical/academic communities

[9] http://en.wikipedia.org/wiki/Semantic_Web
[10] www.biomoby.org
[11] www.opencyc.org
[12] www.freebase.com
[13] www.powerset.com
[14] www.trueknowledge.com
[15] www.wolframalpha.com

and later expanding to all domains of interest. As with other service-oriented applications on the Web, the incorporation of computational knowledge services for scientists will be an important aspect of any research cyberinfrastructure.

The myGrid[16] and myExperiment[17] projects demonstrate the benefits of capturing and then sharing, in a semantically rich way, the definitions of workflows in science. Such workflows effectively document the process by which research-related information is produced and the steps taken toward reaching (or unsuccessfully trying to reach) a conclusion. Imagine the possibilities of expanding this idea to all aspects of our interaction with information. Today, for example, when someone enters "GDP of Brazil vs. Japan" as a query in Wolfram|Alpha, the engine knows how to interpret the input and produce a comparison graph of the GDP (gross domestic product) of the two countries. If the query is "Ford," the engine makes an assumption about its interpretation but also provides alternatives (e.g., "person" if the intended meaning might be Henry Ford or Gerald Rudolph Ford, Jr., vs. "business entity" if the intended meaning might be the Ford Motor Company). The context within which specific information is to be interpreted is important in determining what computational work will be performed. The same ideas could be implemented as part of a global research infrastructure, where Wolfram|Alpha could be one of the many available interoperable services that work together to support researchers.

The research community would indeed benefit greatly from a global infrastructure whose focus is on knowledge sharing and in which all applications and services are built with knowledge exchange and processing at their core. This is not to suggest that there should be yet another attempt to unify and centrally manage all knowledge representation. Scientists will always be better at representing and reasoning over their own domain. However, a research infrastructure should accommodate all domains and provide the necessary glue for information to be cross-linked, correlated, and discovered in a semantically rich manner.

Such an infrastructure must provide the right set of services to not only allow access to semantically rich information but also expose computational services that operate on the world's knowledge. Researchers would be able to ask questions related to their domain of expertise, and a sea of knowledge would immediately be accessible to them. The processes of acquiring and sharing knowledge would be automated, and associated tools (e.g., a word processor that records an author's intended

[16] www.mygrid.org.uk
[17] www.myexperiment.org

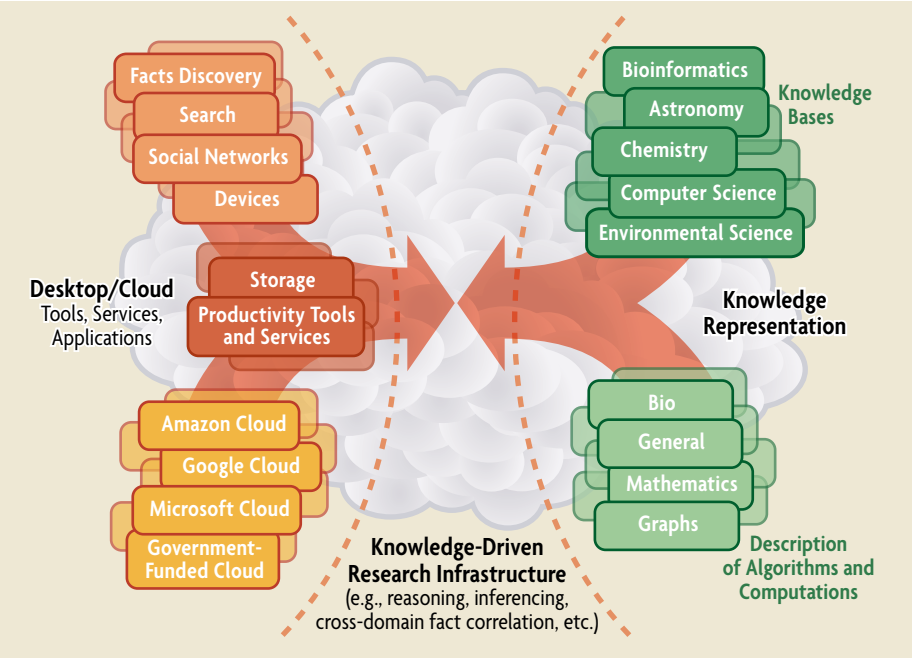


FIGURE 1.
High-level view of a research infrastructure that brings together knowledge bases and computational services.

use of a term[18]) would make it even easier to analyze, do research, and publish results. Natural language processing will aid in the interaction with the knowledge-based ecosystem of information, tools, and services, as shown in Figure 1.

Note that this proposed research infrastructure would not attempt to realize artificial intelligence (AI)—despite the fact that many of the technologies from the Semantic Computing[19] community (from data modeling and knowledge representation to natural language processing and reasoning) have emerged from work in

[18] http://ucsdbiolit.codeplex.com

[19] A distinction is assumed between the general approach of computing based on semantic technologies (machine learning, neural networks, ontologies, inference, etc.) and the Semantic Web as described in [5] and [6], which refers to a specific ecosystem of technologies, such as RDF and OWL. The Semantic Web technologies are considered to be just some of the many tools at our disposal when building semantics-based and knowledge-based solutions.

the AI field over the decades. The primary focus of the proposed cyberinfrastructure is automated knowledge management rather than intelligence.

MASHING UP KNOWLEDGE

Interdisciplinary research has gained a lot of momentum, especially as the result of eScience and cyberinfrastructure activities. Technology has played an enabling role by primarily supporting collaboration, sharing of information, and data management within the context of a research project. In the future, researchers should not have to think about how their questions, assumptions, theories, experiments, or data correlate with existing knowledge across disciplines in one scientific domain or even across domains.

The process of combining information from existing scientific knowledge generated by different researchers at different times and in different locations, including the specific methodologies that were followed to produce conclusions, should be automatic and implicitly supported by the research infrastructure.[20] For example, it should be trivial for a young Ph.D. researcher in chemistry to pose work items to a computer as natural language statements like "Locate 100,000 molecules that are similar to the known HIV protease inhibitors, then compute their electronic properties and dock them into viral escape mutants." This illustrates the use of natural language processing and also the need for researchers to agree on vocabularies for capturing knowledge—something already occurring in many scientific domains through the use of Semantic Web technologies. Furthermore, the example illustrates the need to be able to capture the computational aspects of how existing knowledge is processed and how new facts are generated.

The research community has already started working on bringing the existing building blocks together to realize a future in which machines can further assist researchers in managing and processing knowledge. As an example, the oreChem[21] project aims to automate the process by which chemistry-related knowledge captured in publications is extracted and represented in machine-processable formats, such as the Chemistry Markup Language (CML). Through the use of chemistry-related ontologies, researchers will be able to declaratively describe the computations they would like to perform over the body of machine-processable knowledge.

[20] Assuming that open access to research information has become a reality.
[21] http://research.microsoft.com/orechem

While projects such as oreChem do not attempt to realize a large-scale infrastructure for computable scientific knowledge, they do represent the first investigations toward such a vision. Going forward, the boundaries of domains will become less rigid so that cross-discipline knowledge (computational) mashups can become an important aspect of any semantics-enabled, knowledge-driven research infrastructure. The ability to cross-reference and cross-correlate information, facts, assumptions, and methodologies from different research domains on a global scale will be a great enabler for our future researchers.

A CALL TO ACTION

Today, platforms that offer implementations of the MapReduce computational pattern (e.g., Hadoop and Dryad) make it easy for developers to perform data-intensive computations at scale. In the future, it will be very important to develop equivalent platforms and patterns to support knowledge-related actions such as aggregation, acquisition, inference, reasoning, and information interpretation. We should aim to provide scientists with a cyberinfrastructure on top of which it should be easy to build a large-scale application capable of exploiting the world's computer-represented scientific knowledge.

The interoperable exchange of information, whether representing facts or processes, is vital to successfully sharing knowledge. Communities need to come together—and many of them are already doing so—in order to agree on vocabularies for capturing facts and information specific to their domains of expertise. Research infrastructures of the future will create the necessary links across such vocabularies so that information can be interlinked as part of a global network of facts and processes, as per Tim Berners-Lee's vision for the Semantic Web.

The future research infrastructures, which will be knowledge driven, will look more like Vannevar Bush's memex than today's data-driven computing machines. As Bush said, "Wholly new forms of encyclopedias will appear, ready made with a mesh of associative trails running through them, ready to be dropped into the memex and there amplified." [7] We are not far from that vision today.

ACKNOWLEDGMENTS

The author would like to thank Peter Murray Rust (University of Cambridge) for his explanation of the oreChem project, Evelyne Viegas (Microsoft Research) for insightful discussions and great ideas over the years on all things related to Semantic Computing, and Tony Hey for his continuous support, encouragement, and trust.

REFERENCES

[1] L. Dirks and T. Hey, "The Coming Revolution in Scholarly Communications & Cyberinfrastructure," *CT Watch Q.*, vol. 3, no. 3, 2007.

[2] National Science Foundation, "Cyberinfrastructure Vision for 21st Century Discovery," March 2007.

[3] J. Taylor (n.d.), "UK eScience Programme," retrieved from www.e-science.clrc.ac.uk.

[4] J. Gray and A. Szalay, "eScience - A Transformed Scientific Method," Presentation to the Computer Science and Technology Board of the National Research Council, Jan. 11, 2007, retrieved from http://research.microsoft.com/en-us/um/people/gray/talks/NRC-CSTB_eScience.ppt. (Edited transcript in this volume.)

[5] T. Berners-Lee, J. A. Hendler, and O. Lasilla, "The Semantic Web," *Scientific American*, vol. 284, no. 5, pp. 35–43, May 2001, www.sciam.com/article.cfm?id=the-semantic-web.

[6] N. Shadbolt, W. Hall, and T. Berners-Lee, "The Semantic Web Revisited," *IEEE Intell. Syst.*, vol. 21, no. 3, pp. 96–101, 2006, doi: 10.1109/MIS.2006.62.

[7] V. Bush, "As We May Think," *The Atlantic*, July 1945, doi: 10.3998/3336451.0001.101.

4. SCHOLARLY COMMUNICATION

Introduction

LEE DIRKS | Microsoft Research

JIM GRAY'S PASSION FOR eSCIENCE WAS ADMIRED BY MANY, but few were aware of his deep desire to apply computing to increase the productivity of scholars and accelerate the pace of discovery and innovation for scientific researchers. Several authors in Part 4 of this book knew and worked with Jim. All of the authors not only share his vision but are actively endeavoring to make it a reality.

Lynch introduces how the Fourth Paradigm applies to the field of scholarly communication. His article is organized around a central question: what are the effects of data-intensive science on the scientific record? He goes on to ask: what has become of the scholarly record—an ever-changing, ever-evolving set of data, publications, and related supporting materials of staggering volume? In this new world, not only does the individual scientist benefit (as the end user), but through data-intensive computing we can expect more cross-domain ventures that accelerate discovery, highlight new connections, and suggest unforeseen links that will speed science forward.

Ginsparg delves into the nuts and bolts of the rapid transformation of scholarly publications. He references key examples of cutting-edge work and promising breakthroughs across multiple disciplines. In the process, he notes the siloed nature of the sciences and encourages us to learn from one another and adopt best practices across discipline boundaries. He also provides a helpful

roadmap that outlines an ideal route to a vision he shared with Jim Gray of "community-driven scientific knowledge curation and creation."

Van de Sompel and Lagoze stress that academics have yet to realize the full potential benefits of technology for scholarly communication. The authors make a crucial point that the hardest issues are social or dependent on humans, which means they cannot be easily resolved by new applications and additional silicon. They call for the development of open standards and interoperability protocols to help mitigate this situation.

The issues of sharing scientific data at an international level are addressed by Fitzgerald, Fitzgerald, and Pappalardo. Scientists sometimes encounter the greatest constraints at the national or regional level, which prevent them from participating in the global scientific endeavor. Citing a specific example, the authors appeal for coordination beyond the scientific community and recommend that policymakers work to avoid introducing impediments into the system.

Wilbanks puts a fine point on a common theme throughout this section: in many ways, scientists are often unwittingly responsible for holding back science. Even though, as professionals, we envision, instrument, and execute on innovative scientific endeavors, we do not always actually adopt or fully realize the systems we have put in place. As an amalgamated population of forward-thinking researchers, we often live behind the computational curve. He notes that it is crucial for connectivity to span all scientific fields and for multidisciplinary work and cooperation across domains, in turn, to fuel revolutionary advancements.

Hannay closes the section by highlighting the interconnectedness of our networked world despite lingering social barriers between various scientific fields. He notes that science's gradual shift from a cottage enterprise to a large-scale industry is part of the evolution of how we conduct science. He provides intriguing examples from around the world of research that can point a way to the future of Web-based communication, and he declares that we are living in an awkward age immediately prior to the advent of semantic reality and interconnectedness.

Research is evolving from small, autonomous scholarly guilds to larger, more enlightened, and more interconnected communities of scientists who are increasingly interdependent upon one another to move forward. In undertaking this great endeavor together—as Jim envisioned—we will see science, via computation, advance further and faster than ever before.

Jim Gray's Fourth Paradigm and the Construction of the Scientific Record

CLIFFORD LYNCH
Coalition for Networked Information

In the latter part of his career, Jim Gray led the thinking of a group of scholars who saw the emergence of what they characterized as a fourth paradigm of scientific research. In this essay, I will focus narrowly on the implications of this fourth paradigm, which I will refer to as "data-intensive science" [1], for the nature of scientific communication and the scientific record.

Gray's paradigm joins the classic pair of opposed but mutually supporting scientific paradigms: theory and experimentation. The third paradigm—that of large-scale computational simulation—emerged through the work of John von Neumann and others in the mid-20th century. In a certain sense, Gray's fourth paradigm provides an integrating framework that allows the first three to interact and reinforce each other, much like the traditional scientific cycle in which theory offered predictions that could be experimentally tested, and these experiments identified phenomena that required theoretical explanation. The contributions of simulation to scientific progress, while enormous, fell short of their initial promise (for example, in long-term weather prediction) in part because of the extreme sensitivity of complex systems to initial conditions and chaotic behaviors [2]; this is one example in which simulation, theory, and experiment in the context of massive amounts of data must all work together.

To understand the effects of data-intensive science on the

scientific record,[1] it is first necessary to review the nature of that record, what it is intended to accomplish, and where it has and hasn't succeeded in meeting the needs of the various paradigms and the evolution of science.

To a first approximation, we can think of the modern scientific record, dating from the 17th century and closely tied to the rise of both science and scholarly societies, as comprising an aggregation of independent scientific journals and conference presentations and proceedings, plus the underlying data and other evidence to support the published findings. This record is stored in a highly distributed and, in some parts, highly redundant fashion across a range of libraries, archives, and museums around the globe. The data and evidentiary components have expanded over time: written observational records too voluminous to appear in journals have been stored in scientific archives, and physical evidence held in natural history museums is now joined by a vast array of digital datasets, databases, and data archives of various types, as well as pre-digital observational records (such as photographs) and new collections of biological materials. While scientific monographs and some specialized materials such as patents have long been a limited but important part of the record, "gray literature," notably technical reports and preprints, have assumed greater importance in the 20th century. In recent years, we have seen an explosion of Web sites, blogs, video clips, and other materials (generally quite apart from the traditional publishing process) become a significant part of this record, although the boundaries of these materials and various problems related to their persistent identification, archiving and continued accessibility, vetting, and similar properties have been highly controversial.

The scientific record is intended to do a number of things. First and foremost, it is intended to *communicate* findings, hypotheses, and insights from one person to another, across space and across time. It is intended to organize: to establish common nomenclature and terminology, to connect related work, and to develop disciplines. It is a vehicle for *building up communities* and for a form of large-scale *collaboration* across space and time. It is a means of documenting, managing, and often, ultimately, resolving controversies and disagreements. It can be used to establish *precedence* for ideas and results, and also (through citation and bibliometrics) to offer evidence for the quality and significance of scientific work. The scientific record is intended to be trustworthy, in several senses. In the small and in the near

[1] For brevity and clearest focus, I've limited the discussion here to science. But just as it's clear that eScience is only a special case of eResearch and data-intensive science is a form of data-intensive scholarship, many of the points here should apply, with some adaptation, to the humanities and the social sciences.

term, pre-publication peer review, editorial and authorial reputation, and transparency in reporting results are intended to ensure confidence in the correctness of individual articles. In the broader sense, across spans of time and aggregated collections of materials, findings are validated and errors or deliberate falsifications, particularly important ones, are usually identified and corrected by the community through post-publication discussion or formal review, reproduction, reuse and extension of results, and the placement of an individual publication's results in the broader context of scientific knowledge.

A very central idea that is related simultaneously to trustworthiness and to the ideas of collaboration and building upon the work of others is that of *reproducibility* of scientific results. While this is an ideal that has often been given only reluctant practical support by some scientists who are intent on protecting what they view as proprietary methods, data, or research leads, it is nonetheless what fundamentally distinguishes science from practices such as alchemy. The scientific record—not necessarily a single, self-contained article but a collection of literature and data within the aggregate record, or an article and all of its implicit and explicit "links" in today's terminology—should make enough data available, and contain enough information about methods and practices, that another scientist could reproduce the same results starting from the same data. Indeed, he or she should be able to do additional work that helps to place the initial results in better context, to perturb assumptions and analytic methods, and to see where these changes lead. It is worth noting that the ideal of reproducibility for sophisticated experimental science often becomes problematic over long periods of time: reproducing experimental work may require a considerable amount of tacit knowledge that was part of common scientific practice and the technology base at the time the experiment was first carried out but that may be challenging and time consuming to reproduce many decades later.

How well did the scientific record work during the long dominance of the first two scientific paradigms? In general, pretty well, I believe. The record (and the institutions that created, supported, and curated it) had to evolve in response to two major challenges. The first was mainly in regard to experimental science: as experiments became more complicated, sophisticated, and technologically mediated, and as data became more extensive and less comprehensively reproduced as part of scientific publications, the linkages between evidence and writings became more complex and elusive. In particular, as extended computation (especially mechanically or electromechanically assisted computation carried out by groups of

human "computers") was applied to data, difficulties in reproducibility began to extend far beyond access to data and understanding of methods. The affordances of a scholarly record based on print and physical artifacts offered little relief here; the best that could be done was to develop organized systems of data archives and set some expectations about data deposit or obligations to make data available.

The second evolutionary challenge was the sheer scale of the scientific enterprise. The literature became huge; disciplines and sub-specialties branched and branched again. Tools and practices had to be developed to help manage this scale—specialized journals, citations, indices, review journals and bibliographies, managed vocabularies, and taxonomies in various areas of science. Yet again, given the affordances of the print-based system, all of these innovations seemed to be too little too late, and scale remained a persistent and continually overwhelming problem for scientists.

The introduction of the third paradigm in the middle of the 20th century, along with the simultaneous growth in computational technologies supporting experimental and theoretical sciences, intensified the pressure on the traditional scientific record. Not only did the underlying data continue to grow, but the output of simulations and experiments became large and complex datasets that could only be summarized, rather than fully documented, in traditional publications. Worst of all, software-based computation for simulation and other purposes became an integral part of the question of experimental reproducibility.[2] It's important to recognize how long it really took to reach the point when computer hardware was reasonably trustworthy in carrying out large-scale floating-point computations.[3] (Even today, we are very limited in our ability to produce provably correct large-scale software; we rely on the slow growth of confidence through long and widespread use, preferably in a range of different hardware and platform environments. Documenting complex software configurations as part of the provenance of the products of data-intensive science remains a key research challenge in data curation and scientific workflow structuring.) The better news was that computational technologies began to help with the management of the enormous and growing body of sci-

[2] Actually, the ability to comprehend and reproduce extensive computations became a real issue for theoretical science as well; the 1976 proof of the four-color theorem in graph theory involved exhaustive computer analysis of a very large number of special cases and caused considerable controversy within the mathematical community about whether such a proof was really fully valid. A more recent example would be the proposed proof of the Kepler Conjecture by Thomas Hales.

[3] The IEEE floating-point standard dates back to only 1985. I can personally recall incidents with major mainframe computers back in the 1970s and 1980s in which shipped products had to be revised in the field after significant errors were uncovered in their hardware and/or microcode that could produce incorrect computational results.

entific literature as many of the organizational tools migrated to online databases and information retrieval systems starting in the 1970s and became ubiquitous and broadly affordable by the mid-1990s.

With the arrival of the data-intensive computing paradigm, the scientific record and the supporting system of communication and publication have reached a Janus moment where we are looking both backward and forward. It has become clear that data and software must be integral parts of the record—a set of first-class objects that require systematic management and curation in their own right. We see this reflected in the emphasis on data curation and reuse in the various cyberinfrastructure and eScience programs [3-6]. These datasets and other materials will be interwoven in a complex variety of ways [7] with scientific papers, now finally authored in digital form and beginning to make serious structural use of the new affordances of the digital environment, and at long last bidding a slow farewell to the initial model of electronic scientific journals, which applied digital storage and delivery technologies to articles that were essentially images of printed pages. We will also see tools such as video recordings used to supplement traditional descriptions of experimental methods, and the inclusion of various kinds of two- or three-dimensional visualizations. At some level, one can imagine this as the perfecting of the traditional scientific paper genre, with the capabilities of modern information technology meeting the needs of the four paradigms. The paper becomes a window for a scientist to not only actively understand a scientific result, but also reproduce it or extend it.

However, two other developments are taking hold with unprecedented scale and scope. The first is the development of reference data collections, often independent of specific scientific research even though a great deal of research depends on these collections and many papers make reference to data in these collections. Many of these are created by robotic instrumentation (synoptic sky surveys, large-scale sequencing of microbial populations, combinatorial chemistry); some also introduce human editorial and curatorial work to represent the best current state of knowledge about complex systems (the annotated genome of a given species, a collection of signaling pathways, etc.) and may cite results in the traditional scientific literature to justify or support assertions in the database. These reference collections are an integral part of the scientific record, of course, although we are still struggling with how best to manage issues such as versioning and the fixity of these resources. These data collections are used in very different ways than traditional papers; most often, they are computed upon rather than simply read.

As these reference collections are updated, the updates may trigger new computations, the results of which may lead to new or reassessed scientific results. More and more, at least some kinds of contributions to these reference data collections will be recognized as significant scholarly contributions in their own right. One might think of this as scientists learning to more comprehensively understand the range of opportunities and idioms for contributing to the scholarly record in an era of data and computationally intensive science.

Finally, the scientific record itself is becoming a major object of ongoing computation—a central reference data collection—at least to the extent to which copyright and technical barriers can be overcome to permit this [8]. Data and text mining, inferencing, integration among structured data collections and papers written in human languages (perhaps augmented with semantic markup to help computationally identify references to particular kinds of objects—such as genes, stars, species, chemical compounds, or places, along with their associated properties—with a higher degree of accuracy than would be possible with heuristic textual analysis algorithms), information retrieval, filtering, and clustering all help to address the problems of the ever-growing scale of the scientific record and the ever-increasing scarcity of human attention. They also help exploit the new technologies of data-intensive science to more effectively extract results and hypotheses from the record. We will see very interesting developments, I believe, as researchers use these tools to view the "public" record of science through the lens of various collections of proprietary knowledge (unreleased results, information held by industry for commercial advantage, or even government intelligence).

In the era of data-intensive computing, we are seeing people engage the scientific record in two ways. *In the small,* one or a few articles at a time, human beings read papers as they have for centuries, but with computational tools that allow them to move beyond the paper to engage the underlying science and data much more effectively and to move from paper to paper, or between paper and reference data collection, with great ease, precision, and flexibility. Further, these encounters will integrate with collaborative environments and with tools for annotation, authoring, simulation, and analysis. But now we are also seeing scholars engage the scientific record *in the large,* as a corpus of text and a collection of interlinked data resources, through the use of a wide range of new computational tools. This engagement will identify papers of interest; suggest hypotheses that might be tested through combinations of theoretical, experimental, and simulation investigations; or at times directly produce new data or results. As the balance of engagement

in the large and in the small shifts (today, it is still predominantly in the small, I believe), we will see this change many aspects of scientific culture and scientific publishing practice, probably including views on open access to the scientific literature, the application of various kinds of markup and the choice of authoring tools for scientific papers, and disciplinary norms about data curation, data sharing, and overall data lifecycle. Further, I believe that in the practice of data-intensive science, one set of data will, over time, figure more prominently, persistently, and ubiquitously in scientific work: the scientific record itself.

ACKNOWLEDGMENTS

My thanks to the participants at the April 24, 2009, Buckland-Lynch-Larsen "Friday Seminar" on information access at the University of California, Berkeley, School of Information for a very helpful discussion on a draft of this material.

REFERENCES

[1] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science*, vol. 323, pp. 1297–1298, Mar. 6, 2009, doi: 10.1126/science.1170411.

[2] Freeman Dyson's 2008 Einstein lecture, "Birds and Frogs," *Notices Am. Math. Soc.*, vol. 56, no. 2, pp. 212–224, Feb. 2009, www.ams.org/notices/200902/rtx090200212p.pdf.

[3] National Science Board, "Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century," National Science Foundation, 2005, www.nsf.gov/pubs/2005/nsb0540/start.jsp.

[4] Association of Research Libraries, "To Stand the Test of Time: Long-term Stewardship of Digital Data Sets in Science and Engineering," Association of Research Libraries, 2006. www.arl.org/pp/access/nsfworkshop.shtml.

[5] Various reports available from the National Science Foundation Office of Cyberinfrastructure, www.nsf.gov/dir/index.jsp?org=OCI, including the Cyberinfrastructure Vision document and the Atkins report.

[6] L. Lyon, "Dealing with Data: Roles, Rights, Responsibilities and Relationships," (consultancy report), UKOLN and the Joint Information Systems Committee (JISC), 2006, www.jisc.ac.uk/whatwedo/programmes/programme_digital_repositories/project_dealing_with_data.aspx.

[7] C. A. Lynch, "The Shape of the Scientific Article in the Developing Cyberinfrastructure," *CT Watch*, vol. 3, no. 3, pp. 5–11, Aug. 2007, www.ctwatch.org/quarterly/articles/2007/08/the-shape-of-the-scientific-article-in-the-developing-cyberinfrastructure.

[8] C. A. Lynch, "Open Computation: Beyond Human-Reader-Centric Views of Scholarly Literatures," in Neil Jacobs, Ed., *Open Access: Key Strategic, Technical and Economic Aspects*. Oxford: Chandos Publishing, 2006, pp. 185–193, www.cni.org/staff/cliffpubs//OpenComputation.pdf.

Text in a Data-centric World

PAUL GINSPARG
Cornell University

I FIRST MET JIM GRAY WHEN HE WAS THE MODERATOR of the database subject area of arXiv, part of the expansion into computer science that arXiv initiated in 1998. Soon afterward, he was instrumental in facilitating the full-text harvest of arXiv by large-scale search engines, beginning with Google and followed by Microsoft and Yahoo!—previous robotic crawls of arXiv being overly restricted in the 1990s due to their flooding of the servers with requests. Jim understood the increasing role of text as a form of data, and the need for text to be ingestible and treatable like any other computable object. In 2005, he was involved in both arXiv and PubMed Central and expressed to me his mystification that while the two repositories served similar roles, they seemed to operate in parallel universes, not connecting in any substantive way. His vision was of a world of scholarly resources—text, databases, and any other associated materials—that were seamlessly navigable and interoperable.

Many of the key open questions regarding the technological transformation of scholarly infrastructure were raised well over a decade ago, including the long-term financial model for implementing quality control, the architecture of the article of the future, and how all of the pieces will merge into an interoperable whole. While answers have remained elusive, there is reason to expect significant near-term progress on at least the latter two

questions. In [1], I described how the range of possibilities for large and comprehensive full-text aggregations were just starting to be probed and offered the PubMed Central database as an exemplar of a forward-looking approach. Its full-text XML documents are parsed to permit multiple "related material views" for a given article, with links to genomic, nucleotide, inheritance, gene expression, protein, chemical, taxonomic, and other related databases. This methodology is now beginning to spread, along with more general forms of semantic enhancement: facilitating automated discovery and reasoning, providing links to related documents and data, providing access to actionable data within articles, and permitting integration of data between articles.

A recent example of semantic enhancement by a publisher is the Royal Society of Chemistry's journal *Molecular BioSystems*.[1] Its enhanced HTML highlights terms in the text that are listed in chemical terminology databases and links them to the external database entries. Similarly, it highlights and links terms from gene, sequence, and cell ontologies. This textual markup is implemented by editors with subject-matter expertise, assisted by automated text-mining tools. An example of a fully automated tool for annotation of scientific terms is EMBL Germany's Reflect,[2] which operates as an external service on any Web page or as a browser plug-in. It tags gene, protein, and small molecule names, and the tagged items are linked to the relevant sequence, structure, or interaction databases.

In a further thought experiment, Shotton et al. [2] marked up an article by hand using off-the-shelf technologies to demonstrate a variety of possible semantic enhancements—essentially a minimal set that would likely become commonplace in the near future. In addition to semantic markup of textual terms and live linkages of DOIs and other URLs where feasible, they implemented a reorderable reference list, a document summary including document statistics, a tag cloud of technical terms, tag trees of marked-up named entities grouped by semantic type, citation analysis (within each article), a "Citations in Context" tooltip indicating the type of citation (background, intellectual precedent, refutation, and so on), downloadable spreadsheets for tables and figures, interactive figures, and data fusion with results from other research articles and with contextual online maps. (See Figure 1.) They emphasize the future importance of domain-specific structured digital abstracts—namely, machine-readable metadata that summarize key data and conclusions of articles, including a list of named entities in the article with precise database iden-

[1] www.rsc.org/Publishing/Journals/mb
[2] http://reflect.ws, winner of the recent Elsevier Grand Challenge (www.elseviergrandchallenge.com).

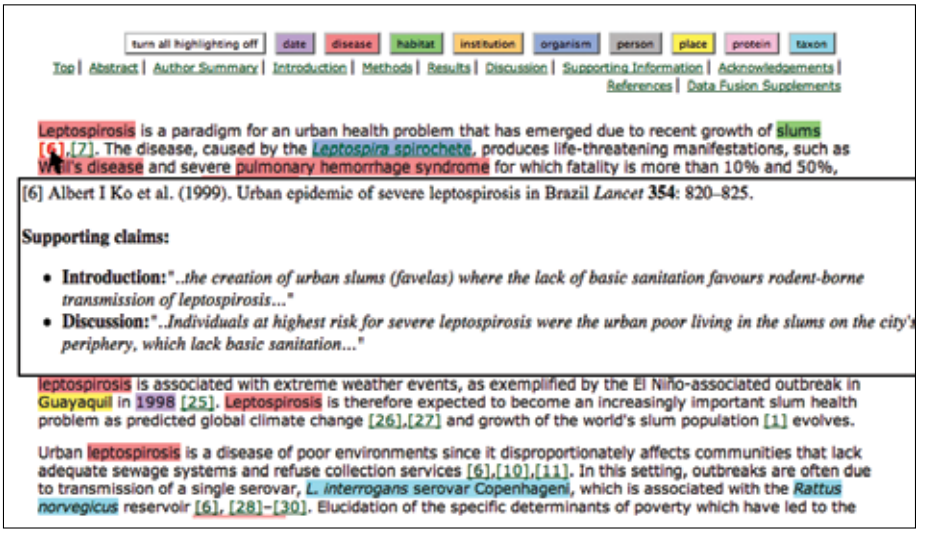


FIGURE 1.

A screenshot of "Exemplar Semantic Enhancements" from http://imageweb.zoo.ox.ac.uk/pub/2008/plospaper/latest, as described in [2]. Different semantic classes of terms are linked and can be optionally highlighted using the buttons in the top row. Hovering the mouse pointer over an in-text reference citation displays a box containing key supporting statements or figures from the cited document.

tifiers, a list of the main results described via controlled vocabulary, and a description, using standard evidence codes, of the methodology employed. The use of controlled vocabularies in this structured summary will enable not only new metrics for article relatedness but also new forms of automated reasoning.

Currently, recognition of named entities (e.g., gene names) in unstructured text is relatively straightforward, but reliable extraction of relationships expressed in conventional text is significantly more difficult. The next generation of automated knowledge extraction and processing tools, operating on structured abstracts and semantically enhanced text, will bring us that much closer to direct searching and browsing of "knowledge"—i.e., via synthesized concepts and their relationships. Further enhancements will include citation network analysis, automated image analysis, more generalized data mashups, and prekeyed or configurable algorithms that provide new types of semantic lenses through which to view the text, data, and images. All of these features can also be federated into hub environments where

users can annotate articles and related information, discover hidden associations, and share new results.

In the near term, semantic text enhancement will be performed by a combination of semi-supervised tools used by authors,[3] tools used by editors, and automated tools applied to both new and archival publications. Many legacy authors will be unwilling to spend time enhancing their documents, especially if much additional effort is required. Certainly many publishers will provide the markup as a value-added component of the publication process—i.e., as part of their financial model. The beneficial effects of this enhancement, visible to all readers, will create pressure in the open sector for equally powerful tools, perhaps after only a small time lag as each new feature is developed. It is more natural to incorporate the semantics from the outset rather than trying to layer it on afterwards—and in either case, PDF will not provide a convenient transport format. With the correct document format, tools, and incentives, authors may ultimately provide much of the structural and semantic metadata during the course of article writing, with marginal additional effort.

In the longer term, there remains the question of where the semantic markup should be hosted, just as with other data published to the Web: Should publishers host datasets relevant to their own publications, or should there be independent SourceForge-like data repositories? And how should the markup be stored: as triple-stores internal to the document or as external attachments specifying relationships and dependencies? As knowledge progresses, there will be new linkages, new things to annotate, and existing annotations that may lead to changed resources or data. Should it be possible to peel these back and view the document in the context of any previous time frame?

To avoid excessive one-off customization, the interactions between documents and data and the fusion of different data sources will require a generic, interoperable semantic layer over the databases. Such structures will also make the data more accessible to generic search engines, via keyword searches and natural-language queries. Having the data accessible in this way should encourage more database maintainers to provide local semantic interfaces, thereby increasing integration into the global data network and amplifying the community benefits of open access to text and data. Tim Berners-Lee[4] has actively promoted the notion of linked data

[3] For example, Pablo Fernicola's "Article Authoring Add-in for Microsoft Office Word 2007," www.microsoft.com/downloads/details.aspx?familyid=09c55527-0759-4d6d-ae02-51e90131997e.

[4] www.w3.org/DesignIssues/LinkedData.html

for all such purposes, not just by academics or for large and commonly used databases. Every user makes a small contribution to the overall structure by linking an object to a URI, which can be dereferenced to find links to more useful data. Such an articulated semantic structure facilitates simpler algorithms acting on World Wide Web text and data and is more feasible in the near term than building a layer of complex artificial intelligence to interpret free-form human ideas using some probabilistic approach.

New forms of interaction with the data layer are also embedded in discussions of Wolfram|Alpha,[5] a new resource (made publicly available only after this writing) that uses substantial personnel resources to curate many thousands of data feeds into a format suitable for manipulation by a Mathematica algorithmic and visualization engine. Supplemented by a front end that interprets semi-natural-language queries, this system and its likely competition will dramatically raise user expectations for new forms of synthesized information that is available directly via generic search engines. These applications will develop that much more quickly over data repositories whose semantic layer is curated locally rather than requiring centralized curation.

Much of the recent progress in integrating data with text via semantic enhancement, as described above, has been with application to the life sciences literature. In principle, text mining and natural-language processing tools that recognize relevant entities and automatically link to domain-specific ontologies have natural analogs in all fields—for example, astronomical objects and experiments in astronomy; mathematical terms and theorems in mathematics; physical objects, terminology, and experiments in physics; and chemical structures and experiments in chemistry. While data-intensive science is certainly the norm in astrophysics, the pieces of the data network for astrophysics do not currently mesh nearly as well as in the life sciences. Most paradoxically, although the physics community was ahead in many of these digital developments going back to the early 1990s (including the development of the World Wide Web itself at CERN, a high-energy physics lab) and in providing open access to its literature, there is currently no coordinated effort to develop semantic structures for most areas of physics. One obstacle is that in many distributed fields of physics, such as condensed-matter physics, there are no dominant laboratories with prominent associated libraries to establish and maintain global resources.

[5] www.wolframalpha.com, based on a private demonstration on April 23, 2009, and a public presentation on April 28, 2009, http://cyber.law.harvard.edu/events/2009/04/wolfram.

In the biological and life sciences, it's also possible that text will decrease in value over the next decade compared with the semantic services that direct researchers to actionable data, help interpret information, and extract knowledge [3]. In most scientific fields, however, the result of research is more than an impartial set of database entries. The scientific article will retain its essential role of using carefully selected data to persuade readers of the truth of its author's hypotheses. Database entries will serve a parallel role of providing access to complete and impartial datasets, both for further exploration and for automated data mining. There are also important differences among areas of science in the role played by data. As one prominent physicist-turned-biologist commented to me recently, "There are no fundamental organizing principles in biology"[6]—suggesting that some fields may always be intrinsically more data driven than theory driven. Science plays different roles within our popular and political culture and hence benefits from differing levels of support. In genomics, for example, we saw the early development of GenBank, its adoption as a government-run resource, and the consequent growth of related databases within the National Library of Medicine, all heavily used.

It has also been suggested that massive data mining, and its attendant ability to tease out and predict trends, could ultimately replace more traditional components of the scientific method [4]. This viewpoint, however, confuses the goals of fundamental theory and phenomenological modeling. Science aims to produce far more than a simple mechanical prediction of correlations; instead, its goal is to employ those regularities extracted from data to construct a unified means of understanding them *a priori*. Predictivity of a theory is thus primarily crucial as a validator of its conceptual content, although it can, of course, have great practical utility as well.

So we should neither overestimate the role of data nor underestimate that of text, and all scientists should track the semantic enhancement of text and related data-driven developments in the biological and life sciences with great interest—and perhaps with envy. Before too long, some archetypal problem might emerge in the physical sciences[7] that formerly required many weeks of complex query traversals of databases, manually maintained browser tabs, impromptu data analysis scripts, and all the rest of the things we do on a daily basis. For example, a future researcher with seamless semantic access to a federation of databases—including band structure properties and calculations, nuclear magnetic resonance (NMR)

[6] Wally Gilbert, dinner on April 27, 2009. His comment may have been intended in a more limited context than implied here.

[7] As emphasized to me by John Wilbanks in a discussion on May 1, 2009.

and X-ray scattering measurements, and mechanical and other properties—might instantly find a small modification to a recently fabricated material to make it the most efficient photovoltaic ever conceived. Possibilities for such progress in finding new sources of energy or forestalling long-term climate change may already be going unnoticed in today's unintegrated text/database world. If classes of such problems emerge and an immediate solution can be found via automated tools acting directly on a semantic layer that provides the communication channels between open text and databases, then other research communities will be bootstrapped into the future, benefiting from the new possibilities for community-driven scientific knowledge curation and creation embodied in the Fourth Paradigm.

REFERENCES

[1] P. Ginsparg, "Next-Generation Implications of Open Access," www.ctwatch.org/quarterly/articles/2007/08/next-generation-implications-of-open-access, accessed Aug. 2007.

[2] D. Shotton, K. Portwin, G. Klyne, and A. Miles, "Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article," *PLoS Comput. Biol.*, vol. 5, no. 4, p. e1000361, 2009, doi: 10.1371/journal.pcbi.1000361.

[3] P. Bourne, "Will a Biological Database Be Different from a Biological Journal?" *PLoS Comput. Biol.*, vol. 1, no. 3, p. e34, 2005, doi: 10.1371/journal.pcbi.0010034. This article was intentionally provocative.

[4] C. Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired*, June 2008, www.wired.com/science/discoveries/magazine/16-07/pb_theory. This article was also intentionally provocative.

All Aboard: Toward a Machine-Friendly Scholarly Communication System

HERBERT VAN DE SOMPEL
Los Alamos National Laboratory

CARL LAGOZE
Cornell University

"The current scholarly communication system is nothing but a scanned copy of the paper-based system."

THIS SENTENCE, WHICH WE USED for effect in numerous conference presentations and eventually fully articulated in a 2004 paper [1], is still by and large true. Although scholarly publishers have adopted new technologies that have made access to scholarly materials significantly easier (such as the Web and PDF documents), these changes have not realized the full potential of the new digital and networked reality. In particular, they do not address three shortcomings of the prevailing scholarly communication system:

- Systemic issues, particularly the unbreakable tie in the publication system between the act of making a scholarly claim and the peer-review process
- Economic strains that are manifested in the "serials crisis," which places tremendous burdens on libraries
- Technical aspects that present barriers to an interoperable information infrastructure

We share these concerns about the state of scholarly communication with many others worldwide. Almost a decade ago, we

collaborated with members of that global community to begin the Open Archives Initiative (OAI), which had a significant impact on the direction and pace of the Open Access movement. The OAI Protocol for Metadata Harvesting (OAI-PMH) and the concurrent OpenURL effort reflected our initial focus on the process-related aspects of scholarly communication. Other members of the community focused on the scholarly content itself. For example, Peter Murray-Rust addressed the flattening of structured, machine-actionable information (such as tabular data and data points underlying graphs) into plain text suited only for human consumption [2].

A decade after our initial work in this area, we are delighted to observe the rapid changes that are occurring in various dimensions of scholarly communication. We will touch upon three areas of change that we feel are significant enough to indicate a fundamental shift.

AUGMENTING THE SCHOLARLY RECORD WITH A MACHINE-ACTIONABLE SUBSTRATE

One motivation for machine readability is the flood of literature that makes it impossible for researchers to keep up with relevant scholarship [3]. Agents that *read* and *filter* on scholars' behalf can offer a solution to this problem. The need for such a mechanism is heightened by the fact that researchers increasingly need to absorb and process literature across disciplines, connecting the dots and combining existing disparate findings to arrive at new insights. This is a major issue in life sciences fields that are characterized by many interconnected disciplines (such as genetics, molecular biology, biochemistry, pharmaceutical chemistry, and organic chemistry). For example, the lack of uniformly structured data across related biomedical domains is cited as a significant barrier to translational research—the transfer of discoveries in basic biological and medical research to application in patient care at the clinical level [4].

Recently, we have witnessed a significant push toward a machine-actionable representation of the knowledge embedded in the life sciences literature, which supports reasoning across disciplinary boundaries. Advanced text analysis techniques are being used to extract entities and entity relations from the existing literature, and shared ontologies have been introduced to achieve uniform knowledge representation. This approach has already led to new discoveries based on information embedded in literature that was previously readable only by humans. Other disciplines have engaged in similar activities, and some initiatives are allowing scholars to start publishing entity and entity-relation information at the time of an article's publication, to avoid the post-processing that is current practice [5].

The launch of the international Concept Web Alliance, whose aim is to provide a global interdisciplinary platform to *discuss, design, and potentially certify solutions for the interoperability and usability of massive, dispersed, and complex data,* indicates that the trend toward a machine-actionable substrate is being taken seriously by both academia and the scholarly information industry. The establishment of a machine-actionable representation of scholarly knowledge can help scholars and learners deal with information abundance. It can allow for new discoveries to be made by reasoning over a body of established knowledge, and it can increase the speed of discovery by helping scholars to avoid redundant research and by revealing promising avenues for new research.

INTEGRATION OF DATASETS INTO THE SCHOLARLY RECORD

Even though data have always been a crucial ingredient in scientific explorations, until recently they were not treated as first-class objects in scholarly communication, as were the research papers that reported on findings extracted from the data. This is rapidly and fundamentally changing. The scientific community is actively discussing and exploring implementation of all core functions of scholarly communication—*registration, certification, awareness, archiving, and rewarding* [1]—for datasets.

For example, the Data Pyramid proposed in [6] clearly indicates how attention to trust *(certification)* and digital preservation *(archiving)* for datasets becomes vital as their application reaches beyond personal use and into the realms of disciplinary communities and society at large. The international efforts aimed at enabling the sharing of research data [7] reflect recognition of the need for an infrastructure to facilitate discovery of shared datasets *(awareness)*. And efforts aimed at defining a standard citation format for datasets [8] take for granted that they are primary scholarly artifacts. These efforts are motivated in part by the belief that researchers should gain credit (be *rewarded)* for the datasets they have compiled and shared. Less than a decade or so ago, these functions of scholarly communication largely applied only to the scholarly literature.

EXPOSURE OF PROCESS AND ITS INTEGRATION INTO THE SCHOLARLY RECORD

Certain aspects of the scholarly communication process have been exposed for a long time. Citations made in publications indicate the use of prior knowledge to generate new insights. In this manner, the scholarly citation graph reveals aspects of scholarly dynamics and is thus actively used as a research focus to detect

connections between disciplines and for trend analysis and prediction. However, interpretation of the scholarly citation graph is often error prone due to imperfect manual or automatic citation extraction approaches and challenging author disambiguation issues. The coverage of citation graph data is also partial (top-ranked journals only or specific disciplines only), and unfortunately the most representative graph (Thomson Reuters) is proprietary.

The citation graph problem is indicative of a broader problem: there is no unambiguous, recorded, and visible trace of the evolution of a scholarly asset through the system, nor is there information about the nature of the evolution. The problem is that relationships, which are known at the moment a scholarly asset goes through a step in a value chain, are lost the moment immediately after, in many cases forever. The actual dynamics of scholarship—the interaction/connection between assets, authors, readers, quality assessments about assets, scholarly research areas, and so on—are extremely hard to recover after the fact. Therefore, it is necessary to establish a layer underlying scholarly communication—a grid for scholarly communication that records and exposes such dynamics, relationships, and interactions.

A solution to this problem is emerging through a number of innovative initiatives that make it possible to publish information about the scholarly process in machine-readable form to the Web, preferably at the moment that events of the above-described type happen and hence, when all required information is available.

Specific to the citation graph case, the Web-oriented citation approach explored by the CLADDIER project demonstrates a mechanism for encoding an accurate, crawlable citation graph on the Web. Several initiatives are aimed at introducing author identifiers [9] that could help establish a less ambiguous citation graph. A graph augmented with citation semantics, such as that proposed by the Citation Typing Ontology effort, would also reveal why an artifact is being cited—an important bit of information that has remained elusive until now [10].

Moving beyond citation data, other efforts to expose the scholarly process include projects that aim to share scholarly usage data (the process of paying attention to scholarly information), such as COUNTER, MESUR, and the bX scholarly recommender service. Collectively, these projects illustrate the broad applicability of this type of process-related information for the purpose of collection development, computation of novel metrics to assess the impact of scholarly artifacts [11], analysis of current research trends [12], and recommender systems. As a result of this work, several projects in Europe are pursuing technical solutions for sharing detailed usage data on the Web.

Another example of process capture is the successful myExperiment effort, which provides a social portal for sharing computational workflow descriptions. Similar efforts in the chemistry community allow the publication and sharing of laboratory notebook information on the Web [13].

We find these efforts particularly inspiring because they allow us to imagine a next logical step, which would be the sharing of provenance data. Provenance data reveal the history of inputs and processing steps involved in the execution of workflows and are a critical aspect of scientific information, both to establish trust in the veracity of the data and to support the reproducibility demanded of all experimental science. Recent work in the computer science community [14] has yielded systems capable of maintaining detailed provenance information within a single environment. We feel that provenance information that describes and interlinks workflows, datasets, and processes is a new kind of process-type metadata that has a key role in network-based and data-intensive science—similar in importance to descriptive metadata, citation data, and usage data in article-based scholarship. Hence, it seems logical that eventually provenance information will be exposed so it can be leveraged by a variety of tools for discovery, analysis, and impact assessment of some core products of new scholarship: workflows, datasets, and processes.

LOOKING FORWARD

As described above, the scholarly record will emerge as the result of the intertwining of traditional and new scholarly artifacts, the development of a machine-actionable scholarly knowledge substrate, and the exposure of meta-information about the scholarly process. These facilities will achieve their full potential only if they are grounded in an appropriate and interoperable cyberinfrastructure that is based on the Web and its associated standards. The Web will not only contribute to the sustainability of the scholarly process, but it will also integrate scholarly debate seamlessly with the broader human debate that takes place on the Web.

We have recently seen an increased Web orientation in the development of approaches to scholarly interoperability. This includes the exploration or active use of uniform resource identifiers (URIs), more specifically HTTP URIs, for the identification of scholarly artifacts, concepts, researchers, and institutions, as well as the use of the XML, RDF, RDFS, OWL, RSS, and Atom formats to support the representation and communication of scholarly information and knowledge. These foundational technologies are increasingly being augmented with community-

specific and community-driven yet compliant specializations. Overall, a picture is beginning to emerge in which all constituents of the new scholarly record (both human and machine-readable) are published on the Web, in a manner that complies with general Web standards and community-specific specializations of those standards. Once published on the Web, they can be accessed, gathered, and mined by both human and machine agents.

Our own work on the OAI Object Reuse & Exchange (OAI-ORE) specifications [15], which define an approach to identifying and describing eScience assets that are aggregations of multiple resources, is an illustration of this emerging Web-centric cyberinfrastructure approach. It builds on core Web technologies and also adheres to the guidelines of the Linked Data effort, which is rapidly emerging as the most widespread manifestation of years of Semantic Web work.

When describing this trend toward the use of common Web approaches for scholarly purposes, we are reminded of Jim Gray, who insisted throughout the preliminary discussions leading to the OAI-ORE work that any solution should leverage common feed technologies—RSS or Atom. Jim was right in indicating that many special-purpose components of the cyberinfrastructure need to be developed to meet the requirements of scholarly communication, and in recognizing that others are readily available as a result of general Web standardization activities.

As we look into the short-term future, we are reminded of one of Jim Gray's well-known quotes: "May all your problems be technical." With this ironic comment, Jim was indicating that behind even the most difficult technical problems lies an even more fundamental problem: assuring the integration of the cyberinfrastructure into human workflows and practices. Without such integration, even the best cyberinfrastructure will fail to gain widespread use. Fortunately, there are indications that we have learned this lesson from experience through the years with other large-scale infrastructure projects such as the Digital Libraries Initiatives. The Sustainable Digital Data Preservation and Access Network Partners (DataNet) program funded by the Office of Cyberinfrastructure at the U.S. National Science Foundation (NSF) has recently awarded funding for two 10-year projects that focus on cyberinfrastructure as a sociotechnical problem—one that requires both knowledge of technology and understanding of how the technology integrates into the communities of use. We believe that this wider focus will be one of the most important factors in changing the nature of scholarship and the ways that it is communicated over the coming decade.

We are confident that the combination of the continued evolution of the

Web, new technologies that leverage its core principles, and an understanding of the way people use technology will serve as the foundation of a fundamentally rethought scholarly communication system that will be friendly to both humans and machines. With the emergence of that system, we will happily refrain from using our once-beloved scanned copy metaphor.

REFERENCES

[1] H. Van de Sompel, S. Payette, J. Erickson, C. Lagoze, and S. Warner, "Rethinking Scholarly Communication: Building the System that Scholars Deserve," *D-Lib Mag.*, vol. 10, no. 9, 2004, www.dlib.org/dlib/september04/vandesompel/09vandesompel.html.

[2] P. Murray-Rust and H. S. Rzepa, "The Next Big Thing: From Hypermedia to Datuments," *J. Digit. Inf.*, vol. 5, no. 1, 2004.

[3] C. L. Palmer, M. H. Cragin, and T. P. Hogan, "Weak information work in scientific discovery," *Inf. Process. Manage.*, vol. 43, no. 3., pp. 808–820, 2007, doi: 10.1016/j.ipm.2006.06.003.

[4] A. Ruttenberg, T. Clark, W. Bug, M. Samwald, O. Bodenreider, H. Chen, D. Doherty, K. Forsberg, Y. Gao, V. Kashyap, J. Kinoshita, J. Luciano, M. S. Marshall, C. Ogbuji, J. Rees, S. Stephens, G. T. Wong, E. Wu, D. Zaccagnini, T. Hongsermeier, E. Neumann, I. Herman, and K. H. Cheung, "Advancing translational research with the Semantic Web," *BMC Bioinf.*, vol. 8, suppl. 3, p. S2, 2007, doi: 10.1186/1471-2105-8-S3-S2.

[5] D. Shotton, K. Portwin, G. Klyne, and A. Miles, "Adventures in Semantic Publishing: Exemplar Semantic Enhancements of a Research Article," *PLoS Comput.* Biol., vol. 5, no. 4, p. e1000361, 2009, doi: 10.1371/journal.pcbi.1000361.

[6] F. Berman, "Got data?: a guide to data preservation in the information age," *Commun. ACM*, vol. 51, no. 12, pp. 50–56, 2008, doi: 10.1145/1409360.1409376.

[7] R. Ruusalepp, "Infrastructure Planning and Data Curation: A Comparative Study of International Approaches to Enabling the Sharing of Research Data," JISC, Nov. 30, 2008, www.dcc.ac.uk/docs/publications/reports/Data_Sharing_Report.pdf.

[8] M. Altman and G. King, "A Proposed Standard for the Scholarly Citation of Quantitative Data," *D-Lib Magazine*, vol. 13, no. 3/4, 2007.

[9] M. Enserink, "Science Publishing: Are You Ready to Become a Number?" *Science*, vol. 323, no. 5922, 2009, doi: 10.1126/science.323.5922.1662.

[10] N. Kaplan, "The norm of citation behavior," *Am. Documentation*, vol. 16. pp. 179–184, 1965.

[11] J. Bollen, H. Van de Sompel, A. Hagberg, and R. Chute, "A Principal Component Analysis of 39 Scientific Impact Measures," *PLoS ONE*, vol. 4, no. 6, p. e6022, 2009, doi: 10.1371/journal.pone.0006022.

[12] J. Bollen, H. Van de Sompel, A. Hagberg, L. Bettencourt, R. Chute, and L. Balakireva, "Clickstream Data Yields High-Resolution Maps of Science," *PLoS ONE*, vol. 4, no. 3, p. e4803, 2009, doi: 10.1371/journal.pone.0004803.

[13] S. J. Coles, J. G. Frey, M. B. Hursthouse, M. E. Light, A. J. Milsted, L. A. Carr, D. De Roure, C. J. Gutteridge, H. R. Mills, K. E. Meacham, M. Surridge, E. Lyon, R. Heery, M. Duke, and M. Day, "An e-science environment for service crystallography from submission to dissemination," *J. Chem. Inf. Model.*, vol. 46, no. 3, 2006, doi: 10.1021/ci050362w.

[14] R. Bose and J. Frew, "Lineage retrieval for scientific data processing: a survey," *ACM Comput. Surv.* (CSUR), vol. 37, no. 1, pp. 1–28, 2005, doi: 10.1145/1057977.1057978.

[15] H. Van de Sompel, C. Lagoze, C. E. Nelson, S. Warner, R. Sanderson, and P. Johnston, "Adding eScience Publications to the Data Web," *Proc. Linked Data on the Web 2009*, Madrid.

The Future of Data Policy

ANNE FITZGERALD
BRIAN FITZGERALD
KYLIE PAPPALARDO
Queensland University of Technology

Advances in information and communication technologies have brought about an information revolution, leading to fundamental changes in the way that information is collected or generated, shared, and distributed [1, 2]. The importance of establishing systems in which research findings can be readily made available to and used by other researchers has long been recognized in international scientific collaborations. Acknowledgment of the need for data access and sharing is most evident in the framework documents underpinning many of the large-scale observational projects that generate vast amounts of data about the Earth, water, the marine environment, and the atmosphere.

For more than 50 years, the foundational documents of major collaborative scientific projects have typically included as a key principle a commitment to ensuring that research outputs will be openly and freely available. While these agreements are often entered into at the international level (whether between governments or their representatives in international organizations), individual researchers and research projects typically operate locally, within a national jurisdiction. If the data access principles adopted by international scientific collaborations are to be effectively implemented, they must be supported by the national policies and laws in place in the countries in which participating researchers

are operating. Failure to establish a bridge between, on the one hand, data access principles enunciated at the international level and, on the other hand, the policies and laws at the national level means that the benefits flowing from data sharing are at risk of being thwarted by domestic objectives [3].

The need for coherence among data sharing principles adopted by international science collaborations and the policy and legal frameworks in place in the national jurisdictions where researchers operate is highlighted by the Global Earth Observation System of Systems[1] (GEOSS) initiated in 2005 by the Group on Earth Observations (GEO) [1, p. 125]. GEOSS seeks to connect the producers of environmental data and decision-support tools with the end users of these products, with the aim of enhancing the relevance of Earth observations to global issues. The end result will be a global public infrastructure that generates comprehensive, near-real-time environmental data, information, and analyses for a wide range of users.

The vision for GEOSS is as a "system of systems," built on existing observational systems and incorporating new systems for Earth observation and modeling that are offered as GEOSS components. This emerging public infrastructure links a diverse and growing array of instruments and systems for monitoring and forecasting changes in the global environment. This system of systems supports policymakers, resource managers, science researchers, and many other experts and decision makers.

INTERNATIONAL POLICIES

One of GEO's earliest actions was to explicitly acknowledge the importance of data sharing in achieving its vision and to agree on a strategic set of data sharing principles for GEOSS [4]:

- There will be full and open exchange of data, metadata and products shared within GEOSS, recognizing relevant international instruments, and national policies and legislation.
- All shared data, metadata, and products will be made available with minimum time delay and at minimum cost.
- All shared data, metadata, and products free of charge or no more than cost of reproduction will be encouraged for research and education.

[1] www.earthobservations.org/index.html

These principles, though significant, are not strictly new. A number of other international policy statements promote public availability and open exchange of data, including the Bermuda Principles (1996) and the Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003) [5].

The Bermuda Principles were developed by scientists involved in the International Human Genome Sequencing Consortium and their funding agencies and represented an agreement among researchers about the need to establish a basis for the rapid and open sharing of prepublication data on gene sequences [6]. The Bermuda Principles required automatic release of sequence assemblies larger than 1 KB and immediate publication of finished annotated sequences. They sought to make the entire gene sequence freely available to the public for research and development in order to maximize benefits to society.

The Berlin Declaration had the goal of supporting the open access paradigm via the Internet and promoting the Internet as a fundamental instrument for a global scientific knowledge base. It defined "open access contribution" to include scientific research results, raw data, and metadata, and it required open access contributions to be deposited in an online repository and made available under a "free, irrevocable, worldwide, right of access to, and a license to copy, use, distribute, transmit and display the work publicly and to make and distribute derivative works, in any digital medium for any responsible purpose, subject to proper attribution of authorship." [7]

In fact, the GEOSS principles map closely to the data sharing principles espoused in the Antarctic Treaty, signed almost 50 years earlier in Washington, D.C., in 1959, which has received sustained attention in Australia, particularly in relation to marine data research.[2] Article III of the Antarctic Treaty states:

> 1. In order to promote international cooperation in scientific investigation in Antarctica, as provided for in Article II of the present Treaty, the Contracting Parties agree that, to the greatest extent feasible and practicable: … (c) scientific observations and results from Antarctica shall be exchanged and made freely available. [8]

The data sharing principles stated in the Antarctic Treaty, the GEOSS 10-Year Implementation Plan, the Bermuda Principles, and the Berlin Declaration, among

[2] Other international treaties with such provisions include the UN Convention on the Law of the Sea, the Ozone Protocol, the Convention on Biodiversity, and the Aarhus Convention.

others, are widely acknowledged to be not only beneficial but crucial to information flows and the availability of data. However, problems arise because, in the absence of a clear policy and legislative framework at the national level, other considerations can operate to frustrate the effective implementation of the data sharing objectives that are central to international science collaborations [5, 9]. Experience has shown that without an unambiguous statement of data access policy and a supporting legislative framework, good intentions are too easily frustrated in practice.

NATIONAL FRAMEWORKS

The key strategy in ensuring that international policies requiring "full and open exchange of data" are effectively acted on in practice lies in the development of a coherent policy and legal framework at a national level. (See Figure 1.) The national framework must support the international principles for data access and sharing but also be clear and practical enough for researchers to follow at a research project level. While national frameworks for data sharing are well established in the United States and Europe, this is not the case in many other jurisdictions (including Australia). Kim Finney of the Antarctic Data Centre has drawn attention to the difficulties in implementing Article III(1)(c) of the Antarctic Treaty in the absence of established data access policies in signatories to the treaty. She points out that being able to achieve the goal set out in the treaty requires a genuine willingness on the part of scientists to make their data available to other researchers. This willingness is lacking, despite the treaty's clear intention that Antarctic science data be "exchanged and made freely available." Finney argues that there is a strong need for a data access policy in Antarctic member states, because without such a policy, the level of conformance with the aspirations set out in the Antarctic Treaty is patchy at best [10] [1, pp. 77–78].

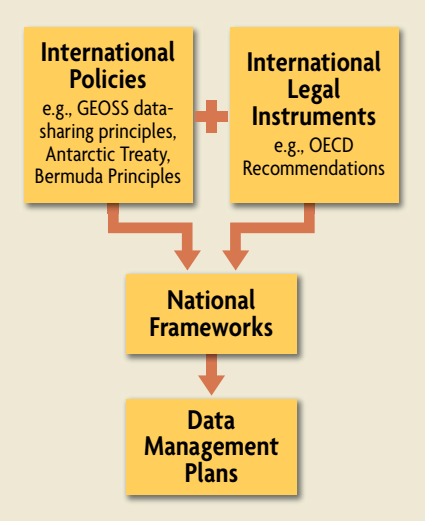


FIGURE 1.
A regulatory framework for data-sharing arrangements.

In the U.S., the Office of Management and Budget (OMB) Circular A-130

establishes the data access and reuse policy framework for the executive branch departments and agencies of the U.S. federal government [11] [1, pp. 174–175]. As well as acknowledging that government information is a valuable public resource and that the nation stands to benefit from the dissemination of government information, OMB Circular A-130 requires that improperly restrictive practices be avoided. Additionally, Circular A-16, entitled "Coordination of Geographic Information and Related Spatial Data Activities," provides that U.S. federal agencies have a responsibility to "[c]ollect, maintain, disseminate, and preserve spatial information such that the resulting data, information, or products can be readily shared with other federal agencies and non-federal users, and promote data integration between all sources." [12] [1, pp. 181–183]

In Europe, the policy framework consists of the broad-reaching Directive on the Re-use of Public Sector Information (2003) (the PSI Directive) [13], as well as the specific directive establishing an Infrastructure for Spatial Information (2007) (the INSPIRE Directive) [14] and the Directive on Public Access to Environmental Information (2003) [15], which obliges public authorities to provide timely access to environmental information.

In negotiating the PSI Directive, the European Parliament and Council of the European Union recognized that the public sector is the largest producer of information in Europe and that substantial social and economic benefits stood to be gained if this information were available for access and reuse. However, European content firms engaging in the aggregation of information resources into value-added information products would be at a competitive disadvantage if they did not have clear policies or uniform practices to guide them in relation to access to and reuse of public sector information. The lack of harmonization of policies and practices regarding public sector information was seen as a barrier to the development of digital products and services based on information obtained from different countries [1, pp. 137–138]. In response, the PSI Directive establishes a framework of rules governing the reuse of existing documents held by the public sector bodies of EU member states. Furthermore, the INSPIRE Directive establishes EU policy and principles relating to spatial data held by or on behalf of public authorities and to the use of spatial data by public authorities in the performance of their public tasks.

Unlike the U.S. and Europe, however, Australia does not currently have a national policy framework addressing access to and use of data. In particular, the current situation with respect to public sector information (PSI) access and reuse is fragmented and lacks a coherent policy foundation, whether viewed in terms of

interactions within or among the different levels of government at the local, state/territory, and federal levels or between the government, academic, and private sectors.[3] In 2008, the "Venturous Australia" report of the Review of the National Innovation System recommended (in Recommendation 7.7) that Australia establish a National Information Strategy to optimize the flow of information in the Australian economy [16]. However, just how a National Information Strategy could be established remains unclear.

A starting point for countries like Australia that have yet to establish national frameworks for the sharing of research outputs has been provided by the Organisation for Economic Co-operation and Development (OECD). At the Seoul Ministerial Meeting on the Future of the Internet Economy in 2008, the OECD Ministers endorsed statements of principle on access to research data produced as a result of public funding and on access to public sector information. These documents establish principles to guide availability of research data, including openness, transparency, legal conformity, interoperability, quality, efficiency, accountability, and sustainability, similar to the principles expressed in the GEOSS statement. The openness principle in the OECD Council's Recommendation on Access to Research Data from Public Funding (2006) states:

> **A) Openness**
> Openness means access on equal terms for the international research community at the lowest possible cost, preferably at no more than the marginal cost of dissemination. Open access to research data from public funding should be easy, timely, user-friendly and preferably Internet-based. [17]

OECD Recommendations are OECD legal instruments that describe standards or objectives that OECD member countries (such as Australia) are expected to implement, although they are not legally binding. However, through long-standing practice of member countries, a Recommendation is considered to have great moral force [2, p. 11]. In Australia, the Prime Minister's Science, Engineering and Innovation Council (PMSEIC) Data for Science Working Group, in its 2006 report "From Data to Wisdom: Pathways to Successful Data Management for Australian Science," recommended that OECD guidelines be taken into account in the development of a strategic framework for management of research data in Australia [18].

The development of a national framework for data management based on

[3] There has been little policy advancement in Australia on the matter of access to government information since the Office of Spatial Data Management's Policy on Spatial Data Access and Pricing in 2001.

principles promoting data access and sharing (such as the OECD Recommendation) would help to incorporate international policy statements and protocols such as the Antarctic Treaty and the GEOSS Principles into domestic law. This would provide stronger guidance (if not a requirement) for researchers to consider and, where practicable, incorporate these data sharing principles into their research project data management plans [5, 9].

CONCLUSION

Establishing data sharing arrangements for complex, international eResearch collaborations requires appropriate national policy and legal frameworks and data management practices. While international science collaborations typically express a commitment to data access and sharing, in the absence of a supporting national policy and legal framework and good data management practices, such objectives are at risk of not being implemented. Many complications are inherent in eResearch science collaborations, particularly where they involve researchers operating in distributed locations. Technology has rendered physical boundaries irrelevant, but legal jurisdictional boundaries remain. If research data is to flow as intended, it will be necessary to ensure that national policies and laws support the data access systems that have long been regarded as central to international science collaborations. In developing policies, laws, and practices at the national level, guidance can be found in the OECD's statements on access to publicly funded research data, the U.S. OMB's Circular A-130, and various EU directives.

It is crucial that countries take responsibility for promoting policy goals for access and reuse of data at all three levels in order to facilitate information flows. It is only by having the proper frameworks in place that we can be sure to keep afloat in the data deluge.

REFERENCES

[1] A. Fitzgerald, "A review of the literature on the legal aspects of open access policy, practices and licensing in Australia and selected jurisdictions," July 2009, Cooperative Research Centre for Spatial Information and Queensland University of Technology, www.aupsi.org.

[2] Submission of the Intellectual Property: Knowledge, Culture and Economy (IP: KCE) Research Program, Queensland University of Technology, to the Digital Economy Future Directions paper, Australian Government, prepared by B. Fitzgerald, A. Fitzgerald, J. Coates, and K. Pappalardo, Mar. 4, 2009, p. 2, www.dbcde.gov.au/__data/assets/pdf_file/0011/112304/Queensland_University_of_Technology_QUT_Law_Faculty.pdf.

[3] B. Fitzgerald, Ed., *Legal Framework for e-Research: Realising the Potential.* Sydney University Press, 2008, http://eprints.qut.edu.au/14439.

[4] Group on Earth Observations (GEO), "GEOSS 10-Year Implementation Plan," adopted Feb. 16,

2005, p. 4, www.earthobservations.org/docs/10-Year%20Implementation%20Plan.pdf.
[5] A. Fitzgerald and K. Pappalardo, "Building the Infrastructure for Data Access and Reuse in Collaborative Research: An Analysis of the Legal Context," OAK Law Project and Legal Framework for e-Research Project, 2007, http://eprints.qut.edu.au/8865.
[6] Bermuda Principles, 1996, www.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml, accessed on June 10, 2009.
[7] Berlin Declaration on Open Access to Knowledge in the Sciences and Humanities (2003), http://oa.mpg.de/openaccess-berlin/berlindeclaration.html, accessed on June 10, 2009.
[8] The Antarctic Treaty (1959), signed in Washington, D.C., Dec. 1, 1959; entry into force for Australia and generally: June 23, 1961, [1961] ATS 12 (Australian Treaty Series, 1961, no. 12), www.austlii.edu.au/cgi-bin/sinodisp/au/other/dfat/treaties/1961/12.html?query=antarctic, accessed June 5, 2009.
[9] A. Fitzgerald, K. Pappalardo, and A. Austin, "Practical Data Management: A Legal and Policy Guide," OAK Law Project and Legal Framework for e-Research Project, 2008, http://eprints.qut.edu.au/14923.
[10] Scientific Committee on Antarctic Research (SCAR) Data and Information Strategy 2008–2013, Joint Committee on Antarctic Data Management (JCADM) and Standing Committee on Antarctic Geographic Information (SC-AGI), authored by K. Finney, Australian Antarctic Data Centre, Australian Antarctic Division (revised May 2008), p. 40, www.jcadm.scar.org/fileadmin/filesystem/jcadm_group/Strategy/SCAR_DIM_StrategyV2-CSKf_final.pdf.
[11] Office of Management and Budget Circular A-130 on Management of Federal Information Resources (OMB Circular A-130), 2000, www.whitehouse.gov/omb/circulars/a130/a130trans4.html.
[12] Office of Management and Budget Circular A-16 on the Coordination of Geographic Information and Related Spatial Data Activities (OMB Circular A-16), issued Jan. 16, 1953, revised 1967, 1990, 2002, Sec. 8, www.whitehouse.gov/omb/circulars_a016_rev/#8.
[13] European Parliament and Council of the European Union, Directive 2003/98/EC of the European Parliament and of the Council of 17 November 2003 on the re-use of the public sector information, 2003, OJ L 345/90, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0098:EN:HTML.
[14] European Parliament and Council of the European Union, Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an infrastructure for spatial information, 2007, OJ L 108/1, Apr. 25, 2007, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:108:0001:01:EN:HTML.
[15] European Parliament and Council of the European Union, Directive 2003/4/EC of the European Parliament and of the Council of 28 January 2003 on public access to environmental information and Repealing Council Directive 90/313/EEC OJL 041, Feb. 14, 2003, pp. 0026–0032, http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:32003L0004:EN:HTML.
[16] Cutler & Company, "Venturous Australia: Building Strength in Innovation," Review of the National Innovation System, p. 95, 2008, www.innovation.gov.au/innovationreview/Pages/home.aspx.
[17] OECD, "Recommendation of the Council concerning Access to Research Data from Public Funding," C(2006)184, Dec. 14, 2006, http://webdomino1.oecd.org/horizontal/oecdacts.nsf/Display/3A5FB1397B5ADFB7C12572980053C9D3?OpenDocument, accessed on June 5, 2009. Note that these have also been published in "OECD Principles and Guidelines for Access to Research Data from Public Funding," 2007.
[18] Prime Minister's Science, Engineering and Innovation Council (PMSEIC) Working Group on Data for Science, "From Data to Wisdom: Pathways to Successful Data Management for Australian Science," Recommendation 9, p. 12, Dec. 2006, www.dest.gov.au/sectors/science_innovation/publications_resources/profiles/Presentation_Data_for_Science.htm.

I Have Seen the Paradigm Shift, and It Is Us

JOHN WILBANKS
Creative Commons

I TEND TO GET NERVOUS WHEN I HEAR TALK OF PARADIGM SHIFTS. The term itself has been debased through inaccurate popular use—even turning into a joke on *The Simpsons*—but its original role in Thomas Kuhn's *Structure of Scientific Revolutions* [1] is worth revisiting as we examine the idea of a Fourth Paradigm and its impact on scholarly communication [2].

Kuhn's model describes a world of science in which a set of ideas becomes dominant and entrenched, creating a worldview (the infamous "paradigm") that itself gains strength and power. This set of ideas becomes powerful because it represents a plausible explanation for observed phenomena. Thus we get the luminiferous aether, the miasma theory of infectious disease, and the idea that the sun revolves around the Earth. The set of ideas, the worldview, the paradigm, gains strength through incrementalism. Each individual scientist tends to work in a manner that adds, bit by bit, to the paradigm. The individual who can make a big addition to the worldview gains authority, research contracts, awards and prizes, and seats on boards of directors.

All involved gain an investment in the set of ideas that goes beyond the ideas themselves. Industries and governments (and the people who work in them) build businesses and policies that depend on the worldview. This adds a layer of defense—an immune system of sorts—that protects the worldview against attack.

Naysayers are marginalized. New ideas lie fallow, unfunded, and unstaffed. Fear, uncertainty, and doubt color perceptions of new ideas, methods, models, and approaches that challenge the established paradigm.

Yet worldviews fall and paradigms shatter when they stop explaining the observed phenomena or when an experiment conclusively proves the paradigm wrong. The aether was conclusively disproven after hundreds of years of incrementalism. As was miasma, as was geocentricism. The time for a shift comes when the old ways of explaining things simply can no longer match the new realities.

This strikes me as being the idea behind Jim Gray's argument about the fourth data paradigm [3] and the framing of the "data deluge"—that our capacity to measure, store, analyze, and visualize data is the new reality to which science must adapt. Data is at the heart of this new paradigm, and it sits alongside empiricism, theory, and simulation, which together form the continuum we think of as the modern scientific method.

But I come to celebrate the first three paradigms, not to bury them. Empiricism and theory got us a long way, from a view of the world that had the sun revolving around the Earth to quantum physics. Simulation is at the core of so much contemporary science, from anthropological re-creations of ancient Rome to weather prediction. The accuracy of simulations and predictions represents the white-hot center of policy debates about economics and climate change. And it's vital to note that empiricism and theory are essential to a good simulation. I can encode a lovely simulation on my screen in which there is no theory of gravity, but if I attempt to drive my car off a cliff, empiricism is going to bite my backside on the way down.

Thus, this is actually not a paradigm shift in the Kuhnian sense. Data is not sweeping away the old reality. Data is simply placing a set of burdens on the methodologies and social habits we use to deal with and communicate our empiricism and our theory, on the robustness and complexity of our simulations, and on the way we expose, transmit, and integrate our knowledge.

What needs to change is our paradigm of ourselves as scientists—not the old paradigms of discovery. When we started to realize that stuff was made of atoms, that we were made of genes, that the Earth revolved around the sun, those were paradigm shifts in the Kuhnian sense. What we're talking about here cuts across those classes of shift. Data-intensive science, if done right, will mean more paradigm shifts of scientific theory, happening faster, because we can rapidly assess our worldview against the "objective reality" we can so powerfully measure.

The data deluge strategy might be better informed by networks than by Kuhnian

dynamics. Networks have a capacity to scale that is useful in our management of the data overload—they can convert massive amounts of information into a good thing so the information is no longer a "problem" that must be "solved." And there is a lesson in the way networks are designed that can help us in exploring the data deluge: if we are to manage the data deluge, we need an open strategy that follows the network experience.

By this I mean the "end-to-end," layer-by-layer, designed information technology and communications networks that are composed of no more than a stack of protocols. The Internet and the Web have been built from documents that propose standard methods for transferring information, describing how to display that information, and assigning names to computers and documents. Because we all agree to use those methods, because those methods can be used by anyone without asking for permission, the network emerges and scales.

In this view, data is not a "fourth paradigm" but a "fourth network layer" (atop Ethernet, TCP/IP, and the Web [4]) that interoperates, top to bottom, with the other layers. I believe this view captures the nature of the scientific method a little better than the concept of the paradigm shift, with its destructive nature. Data is the result of incremental advances in empiricism-serving technology. It informs theory, it drives and validates simulations, and it is served best by two-way, standard communication with those layers of the knowledge network.

To state it baldly, the paradigm that needs destruction is the idea that we as scientists exist as un-networked individuals. Now, if this metaphor is acceptable, it holds two lessons for us as we contemplate network design for scholarly communication at the data-intensive layer.

The first lesson, captured perfectly by David Isenberg, is that the Internet "derives its disruptive quality from a very special property: IT IS PUBLIC." [5] It's public in several ways. The standard specifications that define the Internet are themselves open and public—free to read, download, copy, and make derivatives from. They're open in a copyright sense. Those specifications can be adopted by anyone who wants to make improvements and extensions, but their value comes from the fact that a lot of people use them, not because of private improvements. As Isenberg notes, this allows a set of "miracles" to emerge: the network grows without a master, lets us innovate without asking for permission, and grows and discovers markets (think e-mail, instant messaging, social networks, and even pornography). Changing the public nature of the Internet threatens its very existence. This is not intuitive to those of us raised in a world of rivalrous economic goods and

traditional economic theory. It makes no sense that Wikipedia exists, let alone that it kicks Encyclopedia Britannica to the curb.

As Galileo might have said, however, "And yet it moves." [6] Wikipedia does exist, and the network—a consensual hallucination defined by a set of dry requests for comments—carries Skype video calls for free between me and my family in Brazil. It is an engine for innovation the likes of which we have never seen. And from the network, we can draw the lesson that new layers of the network related to data should encode the idea of publicness—of standards that allow us to work together openly and transfer the network effects we know so well from the giant collection of documents that is the Web to the giant collections of data we can so easily compile.

The second lesson comes from another open world, that of open source software. Software built on the model of distributed, small contributions joined together through technical and legal standardization was another theoretical impossibility subjected to a true Kuhnian paradigm shift by the reality of the Internet. The ubiquitous ability to communicate, combined with the low cost of acquiring programming tools and the visionary application of public copyright licenses, had the strangest impact: it created software that worked, and scaled. The key lesson is that we can harness the power of millions of minds if we standardize, and the products can in many cases outperform those built in traditional, centralized environments. (A good example is the Apache Web server, which has been the most popular Web server software on the Internet since 1996.)

Creative Commons applied these lessons to licensing and created a set of standard licenses for cultural works. These have in turn exploded to cover hundreds of millions of digital objects on the network. Open licensing turns out to have remarkable benefits—it allows for the kind of interoperability (and near-zero transaction costs) that we know from technical networks to occur on a massive scale for rights associated with digital objects such as songs and photographs—and scientific information.

Incentives are the confounding part of all of this to traditional economic theory. Again, this is a place where a Kuhnian paradigm shift is indeed happening—the old theory could not contemplate a world in which people did work for free, but the new reality proves that it happens. Eben Moglen provocatively wrote in 1999 that collaboration on the Internet is akin to electrical induction—an emergent property of the network unrelated to the incentives of any individual contributor. We should not ask why there is an incentive for collaborative software development any more than we ask why electrons move in a current across a wire. We should instead ask,

what is the resistance in the wire, or in the network, to the emergent property? Moglen's Metaphorical Corollaries to Faraday's Law and Ohm's Law[1] still resonate 10 years on.

There is a lot of resistance in the network to a data-intensive layer. And it's actually not based nearly as much on intellectual property issues as it was on software (although the field strength of copyright in resisting the transformation of peer-reviewed literature is very strong and is actively preventing the "Web revolution" in that realm of scholarly communication). With data, problems are caused by copyright,[2] but resistance also comes from many other sources: it's hard to annotate and reuse data, it's hard to send massive data files around, it's hard to combine data that was not generated for recombination, and on and on. Thus, to those who didn't generate it, data has a very short half-life. This resistance originates with the paradigm of ourselves as individual scientists, not the paradigms of empiricism, theory, or simulation.

I therefore propose that our focus be Moglen-inspired and that we resist the resistance. We need investment in annotation and curation, in capacity to store and render data, and in shared visualization and analytics. We need open standards for sharing and exposing data. We need the RFCs (Requests for Comments) of the data layer. And, above all, we need to teach scientists and scholars to work in this new layer of data. As long as we practice a micro-specialization guild culture of training, the social structure of science will continue to provide significant resistance to the data layer.

We need to think of ourselves as connected nodes that need to pass data, test theories, access each others' simulations. And given that every graph about data collection capacity is screaming up exponentially, we need scale in our capacity to use that data, and we need it badly. We need to network ourselves and our knowledge. Nothing else we have designed to date as humans has proven to scale as fast as an open network.

Like all metaphors, the network one has its limits. Networking knowledge is harder than networking documents. Emergent collaboration in software is easier

[1] "Moglen's Metaphorical Corollary to Faraday's Law says that if you wrap the Internet around every person on the planet and spin the planet, software flows in the network. It's an emergent property of connected human minds that they create things for one another's pleasure and to conquer their uneasy sense of being too alone. The only question to ask is, what's the resistance of the network? Moglen's Metaphorical Corollary to Ohm's Law states that the resistance of the network is directly proportional to the field strength of the 'intellectual property' system." [7]

[2] Data receives wildly different copyright treatment across the world, which causes confusion and makes international licensing schemes complex and difficult. [8]

because the tools are cheap and ubiquitous—that's not the case in high-throughput physics or molecular biology. Some of the things that make the Web great don't work so well for science and scholarship because the concept of agreement-based ratings find you only the stuff that represents a boring consensus and not the interesting stuff along the edges.

But there is precious little in terms of alternatives to the network approach. The data deluge is real, and it's not slowing down. We can measure more, faster, than ever before. We can do so in massively parallel fashion. And our brain capacity is pretty well frozen at one brain per person. We have to work together if we're going to keep up, and networks are the best collaborative tool we've ever built as a culture. And that means we need to make our data approach just as open as the protocols that connect computers and documents. It's the only way we can get the level of scale that we need.

There is another nice benefit to this open approach. We have our worldviews and paradigms, our opinions and our arguments. It's our nature to think we're right. But we might be wrong, and we are most definitely not completely right. Encoding our current worldviews in an open system would mean that those who come along later can build on top of us, just as we build on empiricism and theory and simulation, whereas encoding ourselves in a closed system would mean that what we build will have to be destroyed to be improved. An open data layer to the network would be a fine gift to the scientists who follow us into the next paradigm—a grace note of good design that will be remembered as a building block for the next evolution of the scientific method.

REFERENCES

[1] T. S. Kuhn, *The Structure of Scientific Revolutions*. Chicago: University of Chicago Press, 1996.

[2] G. Bell, T. Hey, and A. Szalay, "Beyond the Data Deluge," *Science*, vol. 323, pp. 1297–1298, Mar. 6, 2009, doi: 10.1126/science.1170411.

[3] J. Gray and A. Szalay, "eScience - A Transformed Scientific Method," presentation to the Computer Science and Technology Board of the National Research Council, Mountain View, CA, Jan. 11, 2007. (Edited transcript in this volume.)

[4] Joi Ito, keynote presentation at ETech, San Jose, CA, Mar. 11, 2009.

[5] "Broadband without Internet ain't worth squat," by David Isenberg, keynote address delivered at Broadband Properties Summit, accessed on Apr. 30, 2009, at http://isen.com/blog/2009/04/broadband-without-internet-ain-worth.html.

[6] Wikipedia, http://en.wikipedia.org/wiki/E_pur_si_muove, accessed on Apr. 30, 2009.

[7] E. Moglen, "Anarchism Triumphant: Free Software and the Death of Copyright," *First Monday*, vol. 4, no. 8, Aug. 1999, http://emoglen.law.columbia.edu/my_pubs/nospeech.html.

[8] Science Commons Protocol on Open Access Data, http://sciencecommons.org/projects/publishing/open-access-data-protocol.

From Web 2.0 to the Global Database

TIMO HANNAY
Nature Publishing Group

One of the most articulate of Web commentators, Clay Shirky, put it best. During his "Lessons from Napster" talk at the O'Reilly Peer-to-Peer Conference in 2001, he invited his audience to consider the infamous prediction of IBM's creator, Thomas Watson, that the world market for computers would plateau at somewhere around five [1]. No doubt some of the people listening that day were themselves carrying more than that number of computers on their laps or their wrists and in their pockets or their bags. And that was even before considering all the other computers about them in the room—inside the projector, the sound system, the air conditioners, and so on. But only when the giggling subsided did he land his killer blow. "We now know that that number was wrong," said Shirky. "He overestimated by four." Cue waves of hilarity from the assembled throng.

Shirky's point, of course, was that the defining characteristic of the Web age is not so much the ubiquity of computing devices (transformational though that is) but rather their interconnectedness. We are rapidly reaching a time when any device not connected to the Internet will hardly seem like a computer at all. The network, as they say, is the computer.

This fact—together with the related observation that the dominant computing platform of our time is not Unix or Windows or

Mac OS, but rather the Web itself—led Tim O'Reilly to develop a vision for what he once called an "Internet operating system" [2], which subsequently evolved into a meme now known around the world as "Web 2.0" [3].

Wrapped in that pithy (and now, unfortunately, overexploited) phrase are two important concepts. First, Web 2.0 acted as a reminder that, despite the dot-com crash of 2001, the Web was—and still is—changing the world in profound ways. Second, it incorporated a series of best-practice themes (or "design patterns and business models") for maximizing and capturing this potential. These themes included:

- Network effects and "architectures of participation"
- The Long Tail
- Software as a service
- Peer-to-peer technologies
- Trust systems and emergent data
- Open APIs and mashups
- AJAX
- Tagging and folksonomies
- "Data as the new 'Intel Inside'"

The first of these has widely become seen as the most significant. The Web is more powerful than the platforms that preceded it because it is an open network and lends itself particularly well to applications that enable collaboration. As a result, the most successful Web applications use the network on which they are built to produce their own network effects, sometimes creating apparently unstoppable momentum. This is how a whole new economy can arise in the form of eBay. And how tiny craigslist and Wikipedia can take on the might of mainstream media and reference publishing, and how Google can produce excellent search results by surreptitiously recruiting every creator of a Web link to its cause.

If the Web 2.0 vision emphasizes the global, collaborative nature of this new medium, how is it being put to use in perhaps the most global and collaborative of all human endeavors, scientific research? Perhaps ironically, especially given the origins of the Web at CERN [4], scientists have been relatively slow to embrace

approaches that fully exploit the Web, at least in their professional lives. Blogging, for example, has not taken off in the same way that it has among technologists, political pundits, economists, or even mathematicians. Furthermore, collaborative environments such as OpenWetWare[1] and Nature Network[2] have yet to achieve anything like mainstream status among researchers. Physicists long ago learned to share their findings with one another using the arXiv preprint server,[3] but only because it replicated habits that they had previously pursued by post and then e-mail. Life and Earth scientists, in contrast, have been slower to adopt similar services, such as Nature Precedings.[4]

This is because the barriers to full-scale adoption are not only (or even mainly) technical, but also psychological and social. Old habits die hard, and incentive systems originally created to encourage information sharing through scientific journals can now have the perverse effect of discouraging similar activities by other routes.

Yet even if these new approaches are growing more slowly than some of us would wish, they are still growing. And though the timing of change is difficult to predict, the long-term trends in scientific research are unmistakable: greater specialization, more immediate and open information sharing, a reduction in the size of the "minimum publishable unit," productivity measures that look beyond journal publication records, a blurring of the boundaries between journals and databases, and reinventions of the roles of publishers and editors. Most important of all—and arising from this gradual but inevitable embrace of information technology—we will see an increase in the rate at which new discoveries are made and put to use. Laboratories of the future will indeed hum to the tune of a genuinely new kind of computationally driven, interconnected, Web-enabled science.

Look, for example, at chemistry. That granddaddy of all collaborative sites, Wikipedia,[5] now contains a great deal of high-quality scientific information, much of it provided by scientists themselves. This includes rich, well-organized, and interlinked information about many thousands of chemical compounds. Meanwhile, more specialized resources from both public and private initiatives—notably PubChem[6] and ChemSpider[7]—are growing in content, contributions, and usage

[1] http://openwetware.org
[2] http://network.nature.com
[3] www.arxiv.org
[4] http://precedings.nature.com
[5] http://wikipedia.org
[6] http://pubchem.ncbi.nlm.nih.gov
[7] www.chemspider.com

despite the fact that chemistry has historically been a rather proprietary domain. (Or perhaps in part because of it, but that is a different essay.)

And speaking of proprietary domains, consider drug discovery. InnoCentive,[8] a company spun off from Eli Lilly, has blazed a trail with a model of open, Web-enabled innovation that involves organizations reaching outside their walls to solve research-related challenges. Several other pharmaceutical companies that I have spoken with in recent months have also begun to embrace similar approaches, not principally as acts of goodwill but in order to further their corporate aims, both scientific and commercial.

In industry and academia alike, one of the most important forces driving the adoption of technologically enabled collaboration is sheer necessity. Gone are the days when a lone researcher could make a meaningful contribution to, say, molecular biology without access to the data, skills, or analyses of others. As a result, over the last couple of decades many fields of research, especially in biology, have evolved from a "cottage industry" model (one small research team in a single location doing everything from collecting the data to writing the paper) into a more "industrial" one (large, distributed teams of specialists collaborating across time and space toward a common end).

In the process, they are gathering vast quantities of data, with each stage in the progression being accompanied by volume increases that are not linear but exponential. The sequencing of genes, for example, has long since given way to whole genomes, and now to entire species [5] and ecosystems [6]. Similarly, one-dimensional protein-sequence data has given way to three-dimensional protein structures, and more recently to high-dimensional protein interaction datasets.

This brings changes that are not just quantitative but also qualitative. Chris Anderson has been criticized for his *Wired* article claiming that the accumulation and analysis of such vast quantities of data spells the end of science as we know it [7], but he is surely correct in his milder (but still very significant) claim that there comes a point in this process when "more is different." Just as an information retrieval algorithm like Google's PageRank [8] required the Web to reach a certain scale before it could function at all, so new approaches to scientific discovery will be enabled by the sheer scale of the datasets we are accumulating.

But realizing this value will not be easy. Everyone concerned, not least researchers and publishers, will need to work hard to make the data more useful. This will

[8] www.innocentive.com

involve a range of approaches, from the relatively formal, such as well-defined standard data formats and globally agreed identifiers and ontologies, to looser ones, like free-text tags [9] and HTML microformats [10]. These, alongside automated approaches such as text mining [11], will help to give each piece of information context with respect to all the others. It will also enable two hitherto largely separate domains—the textual, semi-structured world of journals and the numeric, highly structured world of databases—to come together into one integrated whole. As the information held in journals becomes more structured, as that held in many databases becomes more curated, and as these two domains establish richer mutual links, the distinction between them might one day become so fuzzy as to be meaningless.

Improved data structures and richer annotations will be achieved in large part by starting at the source: the laboratory. In certain projects and fields, we already see reagents, experiments, and datasets being organized and managed by sophisticated laboratory information systems. Increasingly, we will also see the researchers' notes move from paper to screen in the form of electronic laboratory notebooks, enabling them to better integrate with the rest of the information being generated. In areas of clinical significance, these will also link to biopsy and patient information. And so, from lab bench to research paper to clinic, from one finding to another, we will join the dots as we explore terra incognita, mapping out detailed relationships where before we had only a few crude lines on an otherwise blank chart.

Scientific knowledge—indeed, all of human knowledge—is fundamentally connected [12], and the associations are every bit as enlightening as the facts themselves. So even as the quantity of data astonishingly balloons before us, we must not overlook an even more significant development that demands our recognition and support: that the information itself is also becoming more interconnected. One link, tag, or ID at a time, the world's data are being joined together into a single seething mass that will give us not just one global computer, but also one global database. As befits this role, it will be vast, messy, inconsistent, and confusing. But it will also be of immeasurable value—and a lasting testament to our species and our age.

REFERENCES

[1] C. Shirky, "Lessons from Napster," talk delivered at the O'Reilly Peer-to-Peer Conference, Feb. 15, 2001, www.openp2p.com/pub/a/p2p/2001/02/15/lessons.html.

[2] T. O'Reilly, "Inventing the Future," 2002, www.oreillynet.com/pub/a/network/2002/04/09/future.html.

[3] T. O'Reilly, "What Is Web 2.0," 2005, www.oreillynet.com/pub/a/oreilly/tim/news/2005/09/30/what-is-web-20.html.
[4] T. Berners-Lee, *Weaving the Web.* San Francisco: HarperOne, 1999.
[5] "International Consortium Announces the 1000 Genomes Project," www.genome.gov/26524516.
[6] J. C. Venter et al., "Environmental genome shotgun sequencing of the Sargasso Sea," *Science,* vol. 304, pp. 66–74, 2004, doi:10.1126/science.1093857.
[7] C. Anderson, "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete," *Wired,* June 2008, www.wired.com/science/discoveries/magazine/16-07/pb_theory.
[8] S. Brin and L. Page, "The Anatomy of a Large-Scale Hypertextual Web Search Engine," 1998, http://ilpubs.stanford.edu:8090/361.
[9] http://en.wikipedia.org/wiki/Tag_(metadata)
[10] http://en.wikipedia.org/wiki/Microformat
[11] http://en.wikipedia.org/wiki/Text_mining
[12] E. O. Wilson, *Consilience: The Unity of Knowledge.* New York: Knopf, 1998.

The Way Forward

CRAIG MUNDIE | Microsoft

THE MULTI-DISCIPLINARY NATURE OF THE ARTICLES collected in this book offers a unique perspective on data-driven scientific discovery—and a glimpse into an exciting future.

As we move into the second decade of the 21st century, we face an extraordinary range of challenges—healthcare, education, energy and the environment, digital access, cyber-security and privacy, public safety, and more. But along with the other contributors to this book, I believe these challenges can be transformed into opportunities with the help of radical new developments in science and technology.

As Jim Gray observed, the first, second, and third paradigms of science—empirical, analytical, and simulation—have successfully carried us to this point in history. Moreover, there is no doubt that if we rely on existing paradigms and technologies, we will continue to make incremental progress. But if we are to achieve *dramatic* breakthroughs, new approaches will be required. We need to embrace the next, fourth paradigm of science.

Jim's vision of this paradigm called for a new scientific methodology focused on the power of *data-intensive science.* Today, that vision is becoming reality. Computing technology, with its pervasive connectivity via the Internet, already underpins almost all scientific study. We are amassing previously unimaginable amounts of data in digital form—data that will help bring about a profound transformation of scientific research and insight. At the same time, computing is on the cusp of a wave of disruptive technological advances—such as multicore architecture,

client-plus-cloud computing, natural user interfaces, and quantum computing—that promises to revolutionize scientific discovery.

Data-intensive science promises breakthroughs across a broad spectrum. As the Earth becomes increasingly instrumented with low-cost, high-bandwidth sensors, we will gain a better understanding of our environment via a virtual, distributed whole-Earth "macroscope." Similarly, the night sky is being brought closer with high-bandwidth, widely available data-visualization systems. This virtuous circle of computing technology and data access will help educate the public about our planet and the Universe at large—making us all participants in the experience of science and raising awareness of its immense benefit to everyone.

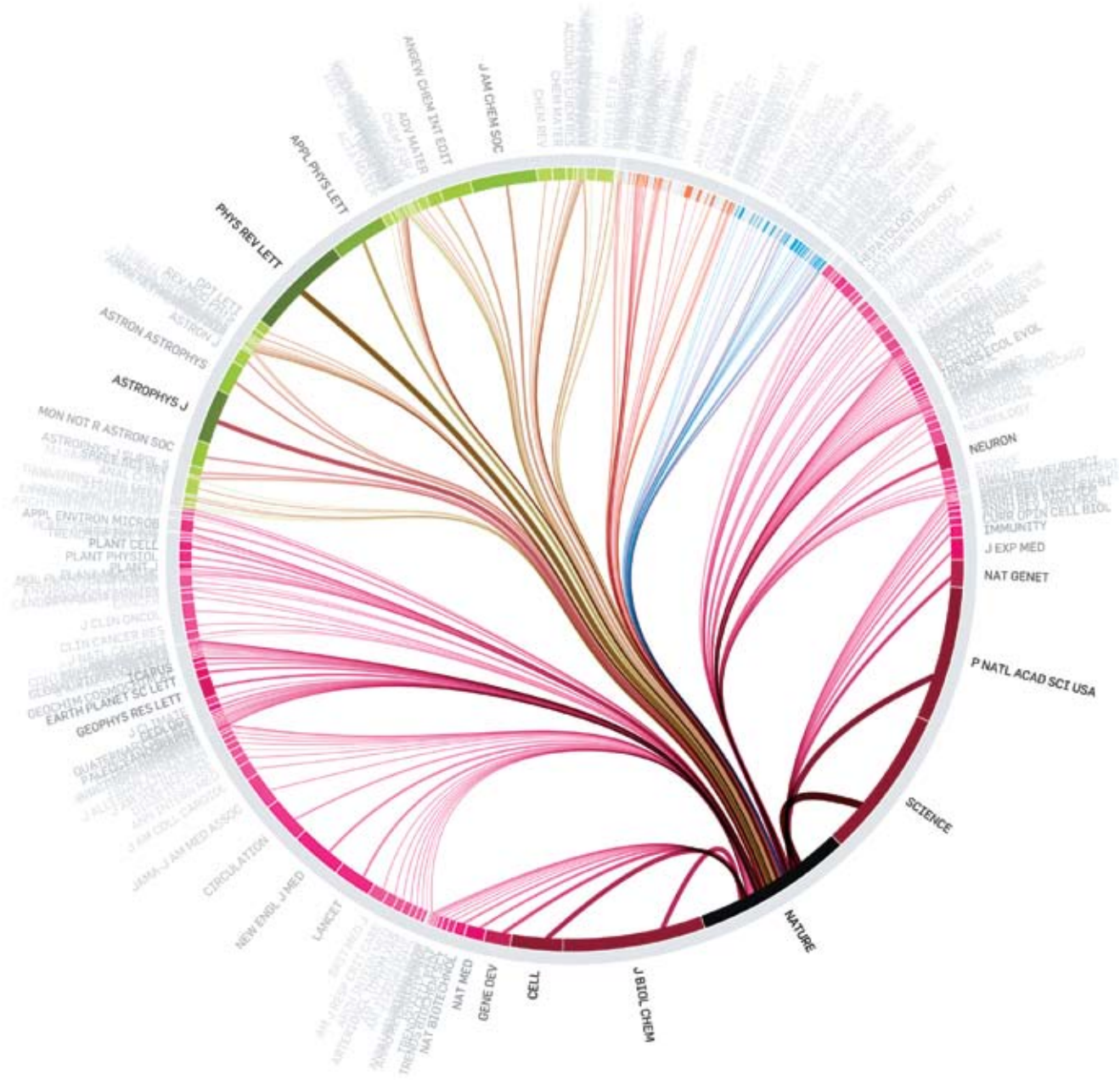
In healthcare, a shift to data-driven medicine will have an equally transformative impact. The ability to compute genomics and proteomics will become feasible on a personal scale, fundamentally changing how medicine is practiced. Medical data will be readily available in real time—tracked, benchmarked, and analyzed against our unique characteristics, ensuring that treatments are as personal as we are individual. Massive-scale data analytics will enable real-time tracking of disease and targeted responses to potential pandemics. Our virtual "macroscope" can now be used on ourselves, as well as on our planet. And all of these advances will help medicine scale to meet the needs of the more than 4 billion people who today lack even basic care.

As computing becomes exponentially more powerful, it will also enable more natural interactions with scientists. Systems that are able to "understand" and have far greater contextual awareness will provide a level of proactive assistance that was previously available only from human helpers. For scientists, this will mean deeper scientific insight, richer discovery, and faster breakthroughs. Another major advance is the emergence of megascale services that are hosted in the cloud and that operate in conjunction with client computers of every kind. Such an infrastructure will enable wholly new data delivery systems for scientists—offering them new ways to visualize, analyze, and interact with their data, which will in turn enable easier collaboration and communication with others.

This enhanced computing infrastructure will make possible the truly global digital library, where the entire lifecycle of academic research—from inception to publication—will take place in an electronic environment and be openly available to all. During the development of scientific ideas and subsequent publishing, scientists will be able to interact virtually with one another—sharing data sources, workflows, and research. Readers, in turn, will be able to navigate the text of a

publication and easily view related presentations, supporting images, video, audio, data, and analytics—all online. Scientific publication will become a 24/7, worldwide, real-time, interactive experience.

I am encouraged to see scientists and computer scientists working together to address the great challenges of our age. Their combined efforts will profoundly and positively affect our future.



The well-formed.eigenfactor project visualizes information flow in science. It came about as a collaboration between the Eigenfactor project (data analysis) and Moritz Stefaner (visualization). This diagram shows the citation links of the journal Nature. *More information and visualizations can be found at http://well-formed.eigenfactor.org.*

Conclusions

TONY HEY, STEWART TANSLEY, AND KRISTIN TOLLE | Microsoft Research

By the mid-1990s, Jim Gray had recognized that the next "big data" challenges for database technology would come from science and not from commerce. He also identified the technical challenges that such data-intensive science would pose for scientists and the key role that IT and computer science could play in enabling future scientific discoveries. The term "eScience" was coined in the year 2000 by John Taylor, when he was director general of the UK Research Councils. Taylor had recognized the increasingly important role that IT must play in the collaborative, multidisciplinary, and data-intensive scientific research of the 21st century and used the term eScience to encompass the collection of tools and technologies needed to support such research. In recognition of the UK eScience initiative, Jim Gray called his research group at Microsoft Research the eScience Group, and he set about working with scientists to understand their problems and learn what tools they needed.

In his talk to the Computer Science and Telecommunications Board of the U.S. National Research Council in 2007, Jim expanded on his vision of data-intensive science and enumerated seven key areas for action by the funding agencies:

1. Foster both the development of software tools and support for these tools.
2. Invest in tools at all levels of the funding pyramid.
3. Foster the development of generic Laboratory Information Management Systems (LIMS).
4. Foster research into scientific data management, data analysis, data visualization, and new algorithms and tools.

5. Establish digital libraries that support other sciences in the same way the National Library of Medicine supports the bio-sciences.
6. Foster the development of new document authoring tools and publication models.
7. Foster the development of digital data libraries that contain scientific data (not just the metadata) and support integration with published literature.

We believe that these challenges to the funding agencies are just as important today. This is why we have introduced this collection of essays, along with a version of Jim's talk to the NRC-CSTB constructed from the transcript of his lecture and his presentation slides. It is also educational to see the continuing momentum and progress of the eScience community since the report "Towards 2020 Science" published by our colleagues at Microsoft Research, Cambridge, UK.[1] That was based on a workshop in July 2005, attended by some of the authors in this new book, and subsequently inspired *Nature's* "2020 Computing" special issue in March 2006.[2]

At the heart of scientific computing in this age of the Fourth Paradigm is a need for scientists and computer scientists to work collaboratively—not in a superior/subordinate relationship, but as equals—with both communities fueling, enabling, and enriching our ability to make discoveries that can bring about productive and positive changes in our world. In this book, we have highlighted healthcare and the environment, just two areas in which humanity faces some of its biggest challenges. To make significant progress, the research community must be supported by an adequate cyberinfrastructure comprising not only the hardware of computing resources, datacenters, and high-speed networks but also software tools and middleware. Jim also envisaged the emergence of a global digital research library containing both the research literature and the research data. Not only are we seeing the maturing of data-intensive science, but we are also in the midst of a revolution in scholarly communication. This is driven not only by technologies such as the Internet, Web 2.0, and semantic annotations but also by the worldwide movement toward open access and open science.

This book is really a labor of love. It started with Jim's desire to enable scientific research through the technologies of computer science—cutting across the disciplines highlighted herein and beyond. We see this book as a continuation of Jim's work with the science community. We deliberately asked our scientific contributors

[1] http://research.microsoft.com/en-us/um/cambridge/projects/towards2020science/background_overview.htm
[2] *Nature*, vol. 440, no. 7083, Mar. 23, 2006, pp. 383–580.

to move out of their professional comfort zones and share their visions for the future of their research fields on a 5-to-10-year horizon. We asked them to write their contributions not only in essay form, which is often a greater challenge than writing a purely technical research article, but often in collaboration with a computer scientist. We are grateful to all of our contributors for rising to this challenge, and we hope that they (and you!) will be pleased with the result.

Several decades ago, science was very discipline-centric. Today, as evidenced by the articles in this book, significant advances are being made as a result of multidisciplinary collaboration—and will continue to be made into the future. The essays in this book present a current snapshot of some of the leading thinking about the exciting partnership between science and computer science—a data revolution—which makes this information timely and potentially fleeting. However, it is our fervent hope and belief that the underlying message presented by the totality of these articles will be durable for many years.

Finally, we offer this book as a call to action for the entire research community, governments, funding agencies, and the public. We urge collaboration toward a common goal of a better life for all humanity. We find ourselves in a phase in which we need to use our scientific understanding to achieve specific goals for the sake of humanity's survival. It is clear that to achieve this aim, we very much need experts with deep scientific knowledge to work closely with those who have deep experience with technology.

This situation is somewhat analogous to the 1940s, when U.S. and European physicists answered an urgent call from governments to collaborate on the Manhattan Project. Today, scientists must collaborate globally to solve the major environmental and health problems facing humanity in a race that is perhaps even more urgent. And ironically, the nuclear physics developed in the Manhattan Project is likely to provide part of the answer in supplying the world with zero-carbon energy.

Tony Hey, Kristin Tolle, and Stewart Tansley

Microsoft External Research, http://research.microsoft.com/collaboration

NEXT STEPS

WE HOPE THIS BOOK WILL INSPIRE YOU to take action as well as embark on further study. We are "walking the talk" ourselves at Microsoft Research. For example, we have reformulated our academic partnership organization, External Research, to focus on the themes presented in this book.

These themes incorporate active research in dynamic fields, so it is hard to track and predict the future evolution of the ideas presented in this book. But here are some suggested ways to remain engaged and to join in the dialogue:

- If you're a scientist, talk to a computer scientist about your challenges, and vice versa.
- If you're a student, take classes in both science and computer science.
- If you're a teacher, mentor, or parent, encourage those in your care toward interdisciplinary study in addition to giving them the option to specialize.
- Engage with the editors and authors of this book through the normal scholarly channels.
- Keep up to date with our eScience research collaborations through our Web site: http://research.microsoft.com.
- Be active in the eScience community—at the Fourth Paradigm Web site below, we suggest helpful resources.

www.fourthparadigm.org

ACKNOWLEDGMENTS

THE EDITORS EXPRESS THEIR HEARTFELT THANKS to all the contributors to this book for sharing their visions within the Fourth Paradigm. We also thank our families and colleagues for their support during the intensive editorial process. The exceptional efforts of the project team, including Ina Chang, Marian Wachter, Celeste Ericsson, and Dean Katz, are also gratefully acknowledged. And, of course, we thank Jim Gray, for inspiring us.

CONTRIBUTORS

Mark R. Abbott
Oregon State University

Dennis D. Baldocchi
University of California, Berkeley

Roger S. Barga
Microsoft Research

Mathias Bavay
WSL Institute for Snow and Avalanche Research SLF

Gordon Bell
Microsoft Research

Chris Bishop
Microsoft Research

José A. Blakeley
Microsoft

Iain Buchan
University of Manchester

Graham Cameron
EMBL-European Bioinformatics Institute

Luca Cardelli
Microsoft Research

Michael F. Cohen
Microsoft Research

Nicholas Dawes
WSL Institute for Snow and Avalanche Research SLF

Del DeHart
Robertson Research Institute

John R. Delaney
University of Washington

David De Roure
University of Southampton

John Dickason
Private practice

Lee Dirks
Microsoft Research

Jeff Dozier
University of California, Santa Barbara

Dan Fay
Microsoft Research

Craig Feied
Microsoft

Anne Fitzgerald
Queensland University of Technology

Brian Fitzgerald
Queensland University of Technology

Peter Fox
Rensselaer Polytechnic Institute

William B. Gail
Microsoft

Dennis Gannon
Microsoft Research

Michael Gillam
Microsoft

Paul Ginsparg
Cornell University

Carole Goble
University of Manchester

Alyssa A. Goodman
Harvard University

Daron Green
Microsoft Research

Jonathan Handler
Microsoft

Timo Hannay
Nature Publishing Group

Charles Hansen
University of Utah

David Heckerman
Microsoft Research

James Hendler
Rensselaer Polytechnic Institute

Eric Horvitz
Microsoft Research

James R. Hunt
University of California, Berkeley, and the Berkeley Water Center

Chris R. Johnson
University of Utah

William Kristan
University of California, San Diego

Carl Lagoze
Cornell University

James Larus
Microsoft Research

Michael Lehning
WSL Institute for Snow and Avalanche Research SLF

Jeff W. Lichtman
Harvard University

Clifford Lynch
Coalition for Networked Information

Simon Mercer
Microsoft Research

Eliza Moody
Microsoft

Craig Mundie
Microsoft

Suman Nath
Microsoft Research

Kylie Pappalardo
Queensland University of Technology

Savas Parastatidis
Microsoft

Marc Parlange
École Polytechnique Fédérale de Lausanne

Valerio Pascucci
University of Utah

Hanspeter Pfister
Harvard University

Catherine Plaisant
University of Maryland

Corrado Priami
Microsoft Research - University of Trento Centre for Computational and Systems Biology and University of Trento

Dan Reed
Microsoft Research

R. Clay Reid
Harvard University

Joel Robertson
Robertson Research Institute

Ben Shneiderman
University of Maryland

Claudio T. Silva
University of Utah

Mark Smith
University of Maryland

Christopher Southan
EMBL-European Bioinformatics Institute

Alexander S. Szalay
The Johns Hopkins University

Kristin Tolle
Microsoft Research

Herbert Van de Sompel
Los Alamos National Laboratory

Catharine van Ingen
Microsoft Research

John Wilbanks
Creative Commons

John Winn
Microsoft Research

Curtis G. Wong
Microsoft Research

Feng Zhao
Microsoft Research

A Few Words About Jim…

Turing award winner and american computer scientist Dr. James Nicholas "Jim" Gray (born 1944, missing at sea on January 28, 2007) was esteemed for his groundbreaking work as a programmer, database expert, engineer, and researcher. He earned his Ph.D. from the University of California, Berkeley, in 1969—becoming the first person to earn a doctorate in computer science at that institution. He worked at several major high-tech companies, including Bell Labs, IBM Research, Tandem, Digital Equipment Corporation, and finally Microsoft Research in Silicon Valley.

Jim joined Microsoft in 1995 as a Senior Researcher, ultimately becoming a Technical Fellow and managing the Bay Area Research Center (BARC). His primary research interests were large databases and transaction processing systems. He had a longstanding interest in scalable computing—building super-servers and work group systems from commodity software and hardware. His work after 2002 focused on eScience: applying computers to solve data-intensive scientific problems. This culminated in his vision (with Alex Szalay) of a "fourth paradigm" of science, a logical progression of earlier, historical phases dominated by experimentation, theory, and simulation.

Jim pioneered database technology and was among the first to develop the technology used in computerized transactions. His work helped develop e-commerce, online ticketing, automated teller machines, and deep databases that enable the success of today's high-quality modern Internet search engines.

In 1998, he received the ACM A.M. Turing Award, the most prestigious honor in computer science, for "seminal contributions to database and transaction process-

ing research and technical leadership in system implementation." He was appointed an IEEE Fellow in 1982 and also received the IEEE Charles Babbage Award.

His later work in database technology has been used by oceanographers, geologists, and astronomers. Among his accomplishments at Microsoft were the TerraServer Web site in collaboration with the U.S. Geological Survey, which paved the way for modern Internet mapping services, and his work on the Sloan Digital Sky Survey in conjunction with the Astrophysical Research Consortium (ARC) and others. Microsoft's WorldWide Telescope software, based on the latter, is dedicated to Jim.

"Jim always reached out in two ways—technically and personally," says David Vaskevitch, Microsoft's senior corporate vice president and chief technical officer in the Platform Technology & Strategy division. "Technically, he was always there first, pointing out how different the future would be than the present."

"Many people in our industry, including me, are deeply indebted to Jim for his intellect, his vision, and his unselfish willingness to be a teacher and a mentor," says Mike Olson, vice president of Embedded Technologies at Oracle Corporation. Adds Shankar Sastry, dean of the College of Engineering at UC Berkeley, "Jim was a true visionary and leader in this field."

"Jim's impact is measured not just in his technical accomplishments, but also in the numbers of people around the world whose work he inspired," says Rick Rashid, senior corporate vice president at Microsoft Research.

Microsoft Chairman Bill Gates sums up Jim's legacy in this way: "The impact of his thinking is continuing to get people to think in a new way about how data and software are redefining what it means to do science."

Such sentiments are frequently heard from the myriad researchers, friends, and colleagues who interacted with Jim over the years, irrespective of their own prominence and reputation. Known, loved, and respected by so many, Jim Gray needs no introduction, so instead we dedicate this book to him and the amazing work that continues in his absence.

—The Editors

GLOSSARY

POWERS OF TEN

exa-	E	1,000,000,000,000,000,000	10^{18}	quintillion
peta-	P	1,000,000,000,000,000	10^{15}	quadrillion
tera-	T	1,000,000,000,000	10^{12}	trillion
giga-	G	1,000,000,000	10^{9}	billion
mega-	M	1,000,000	10^{6}	million
kilo-	k	1,000	10^{3}	thousand
hecto-	h	100	10^{2}	hundred
deca-	da	10	10^{1}	ten
-	-	1	10^{0}	one
deci-	d	0.1	10^{-1}	tenth
centi-	c	0.01	10^{-2}	hundredth
milli-	m	0.001	10^{-3}	thousandth
micro-	μ	0.000001	10^{-6}	millionth
nano-	n	0.000000001	10^{-9}	billionth
pico-	p	0.000000000001	10^{-12}	trillionth

Adapted from http://en.wikipedia.org/wiki/Order_of_magnitude

COMMON ABBREVIATIONS

ASKAP	Australian Square Kilometre Array Pathfinder
ATLUM	Automatic Tape-Collecting Lathe Ultramicrotome
AUV	autonomous underwater vehicle
BPEL	Business Process Execution Language
CCD	charge-coupled device
CEV	Center for Environmental Visualization
CLADDIER	Citation, Location, And Deposition in Discipline and Institutional Repositories
CML	Chemistry Markup Language
CPU	central processing unit
CSTB	Computer Science and Telecommunications Board
DAG	directed acyclic graph
DDBJ	DNA Data Bank of Japan

DOE	Department of Energy
EBI	European Bioinformatics Institute
ECHO	Earth Observing System Clearinghouse
EHR	electronic health record
EMBL	European Molecular Biology Laboratory
EMBL-Bank	European Molecular Biology Laboratory Nucleotide Sequence Database
EOSDIS	Earth Observing System Data and Information System
ET	evapotranspiration
FDA	Food and Drug Administration
FFT	Fast Fourier Transform
FLUXNET	A global network of micrometeorological tower sites
fMRI	functional magnetic resonance imaging
FTP	File Transfer Protocol
GCMD	NASA's Global Change Master Directory
GEOSS	Global Earth Observation System of Systems
GOLD	Genomes OnLine Database
GPU	graphics processing unit
GPGPU	general-purpose graphics processing unit
GUI	graphical user interface
H1N1	swine flu
INSDC	International Nucleotide Sequence Database Collaboration
IT	information technology
KEGG	Kyoto Encyclopedia of Genes and Genomes
KLAS	Keystone Library Automation System
LEAD	Linked Environments for Atmospheric Discovery
LHC	Large Hadron Collider
LIDAR	Light Detection and Ranging
LLNL	Lawrence Livermore National Laboratory
LONI	Laboratory of Neuro Imaging
MESUR	Metrics from Scholarly Usage of Resources
MMI	Marine Metadata Interoperability
NASA	National Aeronautics and Space Administration
NHS	National Health Service (UK)
NIH	National Institutes of Health
NLM	National Library of Medicine

NLM DTD	National Library of Medicine Document Type Definition
NOAA	National Oceanic and Atmospheric Administration
NRC	National Research Council
NSF	National Science Foundation
OAI	Open Archives Initiative
OAI-ORE	Open Archives Initiative Object Reuse and Exchange protocol
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OBO	Open Biomedical Ontologies
OO	object-oriented
OOI	Ocean Observatories Initiative
OWL	Web Ontology Language
Pan-STARRS	Panoramic Survey Telescope And Rapid Response System
PHR	personal health record
PubMed	Free National Library of Medicine online database of biomedical journal articles
RDF	Resource Description Framework
RDFS	RDF Schema
ROV	remotely operated vehicle
RSS	Really Simple Syndication
SCEC	Southern California Earthquake Center
SOA	service-oriented architecture
SWORD	Simple Web-service Offering Repository Deposit
TCP/IP	Transmission Control Protocol/Internet Protocol (the Internet Protocol Suite)
TM	transactional memory
UNICEF	United Nations Children's Fund
UniProt	Universal Protein Resource
URI	Uniform Resource Identifier
USGS	U.S. Geological Survey
VT 100	A Digital Equipment Corporation (DEC) video terminal
WATERS Network	WATer and Environmental Research Systems Network
WHO	World Health Organization
XML	eXtensible Markup Language

INDEX

PHOTO | IMAGE CREDITS

FRONT COVER: Luis Alonso Ocaña/age fotostock. Rights reserved.

PAGE X: *Galileo calculates the magnification of his telescope.* Mary Evans/Photo Researchers, Inc. Rights reserved.

PAGE XVI: *Jim Gray speaking at the Computing in the 21st Century conference in Beijing, October 2006.* Microsoft Research.

PAGE 2: *USGS/NASA image of the Bogda Mountains, China.* U.S. Geological Survey. Public domain.

PAGE 54: *Colored magnetic resonance imaging (MRI) scan of a woman.* Simon Fraser/Photo Researchers, Inc. Rights reserved.

PAGE 108: *A wafer containing the Intel Teraflops Research Chip.* © Intel Corporation. Rights reserved.

PAGE 174: *Central Library, Seattle (Rem Koolhaas, principal architect).* Vetala Hawkins/Filmateria Digital. Rights reserved.

PAGE 222: *Two stars orbit one another in the core of the large emission nebula NGC 6357 in Scorpius, about 8,000 light-years from Earth.* NASA, ESA, and Jesús Maíz Apellániz (Instituto de Astrofísica de Andalucía, Spain). Public domain.

PAGE 226: *Visualization showing the citation links of the journal* Nature. Image courtesy of Moritz Stefaner and Carl Bergstrom, http://well-formed.eigenfactor.org.

PAGE 229: *Tony Hey, Kristin Tolle, and Stewart Tansley of Microsoft External Research.* Vetala Hawkins/Microsoft Corporation.

PAGE 234: *Jim Gray on* Tenacious, *January 2006.* Photo by Tony Hey.

BACK COVER: Microsoft Tag from www.microsoft.com/tag. Get the free app for your phone at http://gettag.mobi and "snap it!"

NOTE: URLs can go offline for various reasons, either temporarily or permanently. Not all of the URLs in this book were still live at the time of publication, but we have successfully accessed such pages using various services such as Internet Archive's Wayback Machine, www.archive.org/web/web.php.

Book design, copyediting, and production by Katz Communications Group, www.katzcommunications.com

Microsoft®

ABOUT THE FOURTH PARADIGM

This book presents the first broad look at the rapidly emerging field of data-intensive science, with the goal of influencing the worldwide scientific and computing research communities and inspiring the next generation of scientists. Increasingly, scientific breakthroughs will be powered by advanced computing capabilities that help researchers manipulate and explore massive datasets. The speed at which any given scientific discipline advances will depend on how well its researchers collaborate with one another, and with technologists, in areas of eScience such as databases, workflow management, visualization, and cloud-computing technologies. This collection of essays expands on the vision of pioneering computer scientist Jim Gray for a new, fourth paradigm of discovery based on data-intensive science and offers insights into how it can be fully realized.

"The impact of Jim Gray's thinking is continuing to get people to think in a new way about how data and software are redefining what it means to do science."

—BILL GATES

"I often tell people working in eScience that they aren't in this field because they are visionaries or super-intelligent—it's because they care about science and they are alive now. It is about technology changing the world, and science taking advantage of it, to do more and do better."

—RHYS FRANCIS, AUSTRALIAN eRESEARCH INFRASTRUCTURE COUNCIL

"One of the greatest challenges for 21st-century science is how we respond to this new era of data-intensive science. This is recognized as a new paradigm beyond experimental and theoretical research and computer simulations of natural phenomena—one that requires new tools, techniques, and ways of working."

—DOUGLAS KELL, UNIVERSITY OF MANCHESTER

"The contributing authors in this volume have done an extraordinary job of helping to refine an understanding of this new paradigm from a variety of disciplinary perspectives."

—GORDON BELL, MICROSOFT RESEARCH

PART NO. 098-115630

ISBN-13: 978-0-9825442-0-4

90000 >

9 780982 544204