



4. SCHOLARLY COMMUNICATION





Introduction

LEE DIRKS | Microsoft Research

JIM GRAY'S PASSION FOR eSCIENCE WAS ADMIRER 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

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).

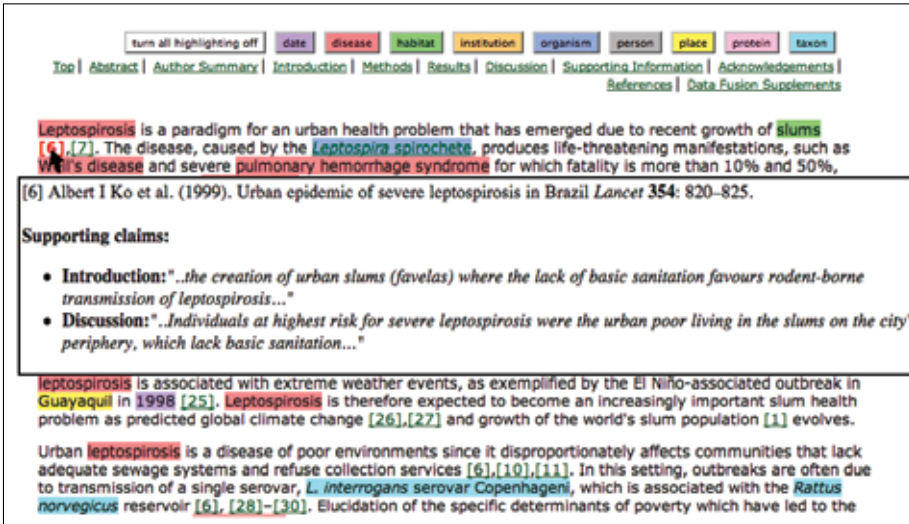


FIGURE 1.

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 triplestores 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-Lee⁴ 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 sciences⁷ 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 Systems¹ (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

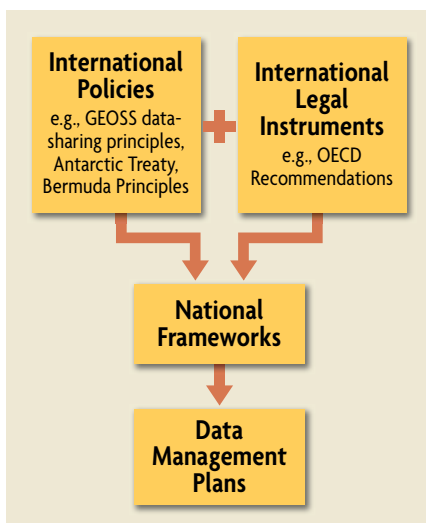


FIGURE 1.
A regulatory framework for data-sharing arrangements.

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].

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/oeecdacts.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

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 geocentrism. 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 Law¹ 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 OpenWetWare¹ and Nature Network² 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 PubChem⁶ and ChemSpider⁷—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_\(metadata\)](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.