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.