◆ *Acceptance Address: Why Us? Why Now?*

Paul Ginsparg
Professor of Physics and Computing and Information Science
Cornell University
Ithaca, New York

I'm honored to be acknowledged by the Council of Science Editors for contributions to the improvement of scientific communication. There's some irony in this, since one interpretation of my little exercise starting 14 years ago was to assess the extent to which editorial disintermediation was possible, by permitting direct scientist-to-scientist communication. I also have to explain that I have only the vaguest inkling of the difficulties facing real editors, coming from my service on the American Physical Society Publications Oversight Committee. But in part, this award testifies to the extraordinary transformative changes in scientific research communication infrastructure over the last decade. We're probably still somewhere in the middle of an ongoing transition, so I'll try to give my own idiosyncratic retrospective perspective on where things stand.

Every generation thinks it's somehow unique, but there are objective reasons to believe that we have been witnessing an essential change in the way science is communicated among research professionals—a fundamental methodologic change that will make the terrain 10 to 20 years from now more different from it was 10 to 20 years ago than in any comparable period. Even the commercial exhibitors at this meeting feature methodologies (Web-based solutions, paperless workflow, XML-based metadata and article ingestion, . . .), most of which didn't exist 10 years ago. At that time, when I was first invited to "scholarly publication meetings", I frequently felt like a visitor from the future, as though showing laser printers to medieval scribes. Now everyone uses the same electronic resources, there's little I can say in this regard that isn't already common knowledge, and happily I'm essentially superfluous. But my experiences can still help illuminate the "Why us? Why now?" questions of what makes this era unique.

I'm part of the first generation to have had computers readily accessible starting in what was then known as junior high school in the late 1960s. Back then, that meant a teletype connected to a remote time-sharing system via a 100-baud acoustically coupled modem and punched paper tape as a storage medium for programs written in Basic and PL/I. This was supplemented in high school by some Fortran programming on punch cards, submitted in batch mode for line printer output the next day. During my freshman year at college in 1973, I first started using e-mail (between Harvard and Stanford on the original Arpanet). As a sign of things to come, my undergraduate class included people like Bill Gates and Steve Ballmer (which probably means that my undergraduate class has the highest average net worth of any undergraduate class ever). I'm also part of the last generation to have experienced the legacy print system, having paid what was then known as a secretary to type my doctoral thesis in 1981. By the mid-1980s, we had switched over entirely to computer typesetting our own articles, typically using TeX (whose underlying text format was also useful to transmit mathematical formulas in informal e-mail communications). By the late 1980s, e-mail connectivity had reached critical mass in my research community of high-energy physics (curiously, e-mail was significantly more useful then than it is now); and in 1987, two collaborators and I first included our e-mail addresses along with physical addresses in an article, initiating that now-universal trend. It was natural by that point to exchange completed manuscripts directly by e-mail and by larger e-mail lists. By 1991, it was natural to organize a more centralized repository and alerting system to facilitate and democratize that exchange, and thus was born *xxx.lanl.gov*, originally an e-mail/ftp server. Finally, it was natural to migrate to the World Wide Web starting in 1993 when usage of that communication protocol started developing critical mass in the research community. Editorial control of the repository was barely necessary back in those antediluvian times, since the Internet was still something of a private playground for academics and we didn't yet need to filter intrusions from the outside world. arXiv was always intended as a forum for communication among research professionals, not as a mechanism for outsiders to communicate to that community.

In the late 1980s, there was little indication that conventional journals would be available in any convenient online format any time soon. Surveys of librarians just a decade ago suggested an expectation of any time from 25 years to a century (in other words, "I don't know but after I'm retired") before we'd be at the point we've already reached in 2005. By the late 1990s, a wide array of major journals had established a significant online presence. Now that I'm back at a university, it's entertaining to see successive generations of students having increasingly adopted the attitude that "if it isn't online, then it may as well not exist." Libraries are recognized as places to buy coffee and to connect to the Internet for instant messaging and search the Web for online information.

The very ability to browse information in this way is a reflection of one of the major lessons of the last decade: the surprising efficacy of brute force computation. Google, for example, came along in 1998 with a relatively simple set of heuristics tied to a creative hardware implementation and provided an unexpectedly powerful methodology for nonexperts to navigate the information in billions of Web pages. It might not have seemed plausible a decade ago that simple search engine methodology could scale to provide systematically useful information from such an ever-increasing set of resources; but search engines are cur-

Acceptance continued

rently essential, certainly do work in a variety of useful ways, and continue to improve. Another example of the power of simple heuristics tied to ample computational power operating on large datasets is provided by aggregated scholarly collections. The American Physical Society's PROLA contains the scanned and OCR'd full text of all its journals back to their inception in 1893, the JSTOR project includes back issues of mainly nonscience journals, and the Astrophysical Data System is a comprehensive aggregation of back issues of nearly all astrophysics journals. The result of aggregation in each case, combined with ease of navigation—via metadata and full-text searches, the ability to follow citation linkages, and links to related resources—results in a resource ultimately far more powerful than might be imagined from the simple sum of its parts. The result is not only greater ease of research, but also improved scholarship, and increased latency of archival reference usage.

But we're still just scratching the surface of what can be done with large and comprehensive full-text aggregations. A forward-looking example of the sort of automated editorial annotation we can expect more generally in the future is given by the PubMed Central database, operated in conjunction with GenBank and other biologic databases at the US National Library of Medicine. In this case, full-text documents are parsed to permit multiple different "views". GenBank accession numbers are recognized in articles referring to sequence data and linked directly to the relevant records in the genomic databases. Protein names are recognized and their appearances in articles linked automatically to the protein and protein-interaction databases. Names of organisms are recognized and linked directly to the taxonomic databases, which are then used to compute a minimal spanning tree of all the organisms contained in a given document. In yet another "view", technical terms are recognized and linked directly to the glossary items in the relevant standard biology or biochemistry textbook in the books database. The enormously powerful sorts of data-mining and number-crunching, already taken for granted as applied to the open-access genomics databases, can be applied to the full text of the entirety of the biology and life-sciences literature and will have just as great a transformative effect on the research done with it.

A decade ago, in the initial euphoria about the promise of electronic scholarly publishing, there was some confusion about the actual costs of publishing due to the paper format. Mistaken claims that as much as 90% of the cost could be eliminated by eliminating paper, combined with hopes that all scholarly output could quickly be made freely accessible, suggested some rapid transformation to a new economic model for supporting editorial costs. In reality, the scholarly publishing industry is far from monolithic and involves large hierarchies in costs (and even larger hierarchies in revenues), but costs associated with producing and disseminating paper were never remotely the dominant component of the publishing costs. There's no time here to describe in any detail either the enormous benefits of open access or the difficulties proponents may have in establishing a sustainable financial model for its fully featured incarnation. Quality control costs real money, but there may yet prove to be a more efficient methodology for providing the same ultimate functionality as the current system, and ultimately that method may be discipline-specific. But a form of open access also appears to be happening by a backdoor route regardless: An article in last week's [14 May 2005] *BMJ* reports that over a third of the high-impact journal articles in a sample of biologic-medical journals published in 2003 could be found at nonjournal Web sites. Given the likelihood that this percentage will only increase over time and given the aforementioned ease of locating these materials, this phenomenon is something with which journals will have to come to grips, and without alienating their contributing authors.

The *arXiv.org* site currently ingests and disseminates roughly 50,000 new articles per year, the vast majority of which are subject to some form of review, whether by journals, conference organizers, or thesis committees. Physics and astronomy journals have learned to take active advantage of the prior availability of the materials, and the resulting symbiotic relation might not have been anticipated 14 years ago. For reasons touched upon above, however, the current situation may not be stable for the long term, and different forms of accommodation may need to evolve. arXiv journals, search engines, and other third-party distribution systems and aggregators all work in their own ways to satisfy readers. From the outset, *arXiv.org* relied on a variety of heuristic screening mechanisms to ensure insofar as possible that submissions are at least "of refereeable quality". That means they satisfy the minimal criterion that they would not be peremptorily rejected by any competent journal editor as nutty, offensive, or otherwise manifestly inappropriate and would instead at least in principle be suitable for review. These mechanisms are an important—if not essential—component of why readers find the arXiv site so useful. Though the most recently submitted articles have not yet necessarily undergone formal review, the vast majority of the articles can, would, or do eventually satisfy editorial requirements somewhere. The flux of submissions intercepted by automated editorial processes has continued to increase since the discovery of the Internet by the rest of the world in the mid-1990s.

One final anecdote before closing: In early 1994, I happened to serve on a committee advising the American Physical Society about putting *Physical Review Letters* online. I pointed to the Web interface I'd set up and suggested this might be a good way for APS to disseminate its documents. A response came back from another committee member along these lines: "Installing and learning to use a World Wide Web browser is a complicated and difficult task—we can't possibly expect this of the average physicist." So they went with a different (and short-lived) platform. This was before Marty Blume arrived as editor-in-chief in 1997, at which point APS began to lead the way, so I now pass the microphone to him.

Editor's Note: Science Editor *hopes to include in a future issue a piece based on the acceptance address by Martin Blume, who spoke on Albert Einstein and peer review.*