SCHOLARLY INFORMATION ARCHITECTURE, 1989-2015

P Ginsparg

Depts of Physics and Computing & Information Science, Cornell University, Ithaca, NY 14853, USA
Email: ginsparg@cornell.edu

ABSTRACT

If we were to start from scratch today to design a quality-controlled archive and distribution system for scientific and technical information, it could take a very different form from what has evolved in the past decade from pre-existing print infrastructure. Ultimately, we might expect some form of global knowledge network for research communications. Over the next decade, there are many technical and non-technical issues to address along the way, everything from identifying optimal formats and protocols for rendering, indexing, linking, querying, accessing, mining, and transmitting the information, to identifying sociological, legal, financial, and political obstacles to realization of ideal systems. What near-term advances can we expect in automated classification systems, authoring tools, and next-generation document formats to facilitate efficient datamining and long-term archival stability? How will the information be authenticated and quality controlled? What differences should be expected in the realization of these systems for different scientific research fields? Can recent technological advances provide not only more efficient means of accessing and navigating the information, but also more cost-effective means of authentication and quality control? Relevant experiences from open electronic distribution of research materials in physics and related disciplines during the past decade are used to illuminate these questions, and some of their implications for proposals to improve the implementation of peer review are then discussed.

Keywords: arXiv, peer review, scholarly publishing, scientific publishing

1 INTRODUCTION

This discussion is based on a keynote presentation given at the CODATA 2002 conference during the session “CODATA 2015” on 1 Oct 2002 (Ginsparg, 2002a). I was asked to help anticipate the sorts of data issues with which CODATA will likely be dealing in the year 2015. I was told 2015 was chosen to be far enough off in the distance to permit creative fantasizing, but not so far as to be impossible to make useful conjectures. My past CODATA experience (Berry, Alexander, Allen, Baumgardner, Branscomb, Cizewski, et al., 1997) suggests that such issues will involve not only technological considerations, but also a highly non-trivial mix of sociological, legal, financial, and political considerations.

In what follows, I will begin with some general considerations about the dramatic technological changes of the recent past. The remainder here will focus not on the full spectrum of data issues, but rather on the subset related specifically to scholarly research infrastructure, the part with which I’m most familiar. Extrapolation of how dissemination of scientific ideas and knowledge will evolve in the near and not–so–near future will touch on many of the issues that will affect data more generally. Berry et al., (1997) argued that the full text of scientific articles falls within CODATA purview as essential data.

2 BACK TO THE PAST

2015 is 13 years in the future, so as a warmup for this exercise, it’s useful to do a thought experiment, and try to recall what a similar talk given 13 years ago would have been able to predict for 2002. In 1989 I had a Toshiba laptop with an 80C86 processor (4.77 MHz, 640K RAM) running DOS for home and travel usage. At work, I shared a department VAX and was permitted a disk allocation of 2000 blocks (equivalent to 1Mb) for all my e-mail and work files. From remote, I could connect to my work machine either with my laptop and external 2400 baud modem, or via the old DECNET or more recently implemented internet. There already existed incipient forms of all the hardware we now take for granted, albeit somewhat slower and with significantly less diskspace and memory.

For software, I was using TeX for all of my technical typesetting requirements, and that program exists in essentially the same form and remains in widespread use for producing documents with technical typesetting requirements. Since the early 1980's, I had been using e-mail for increasing amounts of my professional communication, and my e-mail usage now is little different from what it was 13 years ago, except that it's less fun. It had started as a privileged communication channel with people I knew well, and now it's much more a broadcast channel (not only due to unsolicited spam and e-mail viruses). I probably would not have predicted 13 years ago that certain communications can be more efficient and less ambiguous by phone: a substantive e-mail can take a half hour, and that means only roughly 15 per workday (ignoring all other responsibilities).

When we think of on-line resources, the major development that few of us had foreseen in 1989 was the transformative effect of the WorldWideWeb. The web is a straightforward set of protocols which, combined with a new generation of graphical browsers, altered the expectations of the network experience, and accelerated adoption of network usage by the non-academic community. Few of us had foreseen the extent of the entry of the commercial sector starting in the mid 1990's, the ubiquitous appearance of uniform resource locators (URLs) in advertising, the .com bubble, and the penetration of e-mail and other forms of internet usage into the daily life of an increasing percentage of the population. Many networked research resources, such as SLAC-SPIRES for High Energy Physics, or Med-Line and Genbank for the Biological and Life Sciences, were already very popular prior to the web, but their increased availability, ease of usage, and comprehensiveness, all mediated by the web, led them to achieve a new level of utility, and they've since been joined by a vast array of other high quality resources designed by, and intended for, research professionals.

One lesson from the above considerations is that at least many parts of the computing and networking landscape in 2015 will seem familiar, consisting of enhanced versions of current utilities or ideas. But another lesson of the past couple decades — the surprising efficacy of brute force computing (Arms, 2000) — suggests there will also be some unexpectedly powerful applications, beyond easy anticipation. "Moore's Law" is the phenomenological observation that computer circuit density, hence computing power, doubles roughly every 1.5 years. If that trend continues, it will mean an increase in computing speed by a factor of 400 times in the next 13 years (consistent with an increase in the past 13 years from order 5Mhz to order 2Ghz in personal computer speeds), for an overall increase by a factor of 400*400 = *160,000* from 1989 to 2015. Simple algorithms plus immense computing power applied to large datasets can outperform human intelligence on many useful tasks, and this substitution for human intelligence can be achieved without having to realize the original objectives of the Artificial Intelligence program. Deep Blue's programmers, for example, did not have to understand or emulate Kasparov's thought processes or analytic abilities to win their chess competition. A web-related example is that the search engine methodology of the mid 1990's appeared to be a dead end: it only sort of worked because the amount of information on the web was so limited, but once the web became more comprehensive then every query would return so many results it would become increasingly impossible to find the ones desired. But Google came along in 1998 with a relatively simple set of heuristics tied to an extraordinary hardware implementation, and provided an unexpectedly powerful methodology for non-experts to navigate the information in over 3 billion web pages.

Returning now to the question of scholarly information, in the late 1980's there was little indication that conventional journals would be available in any convenient on-line format any time soon, but by the late 1990s most major journals had established a significant on-line presence. Successive generations of students have increasingly adopted the attitude that "if it isn't on-line, then it may as well not exist." Something qualitative frequently happens at critical threshholds, known colloquially as "tipping points." People formerly accustomed to regular library usage first move to a mix of library and on-line desktop usage, but eventually enough of what they need is on-line that they abandon library usage altogether, forcibly ignoring those materials available only in paper as no longer worth the effort for so small a percentage of potential research materials. Moreover, recent generations of undergraduates have increasingly come to visualize campus libraries as much as a place to buy expresso and connect laptops to a wireless network as a place to find archival resources, and they do not acquire a conventional library habit in the first place.

Prior to the web, access in the U.S. to scientific, medical, and legal information was excellent for those belonging to rich organizations (e.g, a major universities), but was otherwise poor. In many countries of the world, it was poor for everybody. The web has provided access to a significant segment of this information to many throughout the world who did not formerly have access, but now have an internet connection. Moreover, the *percentage* of the information that is freely available continues to increase. We can certainly now foresee and expect on a timescale of decades to have on-line all back issues going back centuries of all or most journals in all major disciplines (whether freely available or not), with the obstacles to achieving this goal more logistical and legal in nature than technical or financial. The resources already available on-line have transformed the way that research is performed, and have improved scholarship. Though there is a wealth of information available, the tools we use to navigate it remain rudimentary and in many cases generic. A next generation set of tools explicitly designed for research professionals to navigate the full breadth of their discipline, plus relevant portions of neighboring disciplines, would have a doubly transformative effect.

The first underpinnings for this next stage will be provided by a next-generation of document formats that transparently expose their metadata, provide for contextual searching, and simultaneously provide a stable archival and renderable format. (A first step in this direction is the "Archiving and Interchange DTD" (NCBI DTD, n.d.), developed in conjunction with the NIH's PubMed Central project (PubMed Central, n.d.).) The next set of advances in natural language processing should permit supplementing the current rudimentary information retrieval methodology ("bag of words") with more sophisticated techniques for determining semantic content and context.

The WorldWideWeb, rendering resources easier to find through linkages, and thereby encouraging the creation of more resources, is the most visible of the qualitative on-line changes of the past 13 years. Extrapolating from its recent past history thus provides a useful place to start constructing forward projections. Looking back a decade, we begin to hit an edge effect, but metaphorically the web in its earliest days was much smaller than the internet in which it resided, and consisted primarily of large well-connected academic institutions. Overall it was small enough that Tim Berners-Lee, its creator, could maintain at its CERN birthplace a list of "all web servers in the world", sorted by continent. Fast-forwarding 5 years to the mid-late '90's, we saw significant further expansion into the academic community and government, more significantly the explosion of sites in the .com sector, and parenthetically the end of the fun era (quirky personal pages, webcams of coffee pots, . . . , losing novelty and perhaps too labor-intensive to maintain for people simultaneously working jobs with actual responsibilities). Finding information was still hit or miss, relatively labor-intensive, and there was no expectation that things should be available.

The web has since undergone a subtle transformation, slowly so perhaps not as noticeable, from what was roughly speaking a static "library" metaphor to a dynamic "breaking news" metaphor. In conjunction with navigational tools such as provided by google.com, the early 21st century web has expanded dramatically from the .com network of the late '90s, and with that bubble burst now encompasses much more of society at large. It covers more public and private institutions, facilitates more commerce, incorporates more newspapers/magazines, and, with the advent of bloggers, more private commentary. Many types of information are virtually guaranteed to be there, somewhere. This includes in particular information regarding a certain class of people: public figures in some generalized sense, including academics, invariably register some web presence. An increasing amount, though perhaps not percentage, of information is available only by subscription. Coverage in parts of the world outside of the U.S. remains spotty, however, particularly in less developed countries and totalitarian regimes. It's likely we're still at the leading cusp of this ongoing transformation, with the advent of bloggers and global news resources just starting to come into their own and driving "Web Phase II".

Extrapolating unimaginatively from the above, we would envision a web in 2015 which manifests some presence from every institution: academic, government, financial, political, social, medical, . . . ; from local and national to global levels. This presence would include representation from every department within those institutions, from every person within those departments, and also provide some record of every public event (seminars, meetings) that occurs within or affects those institutions. (Additional individual footprints might be left by other means such as mobile devices, but we assume that few of those will leave public traces for harvesting.) What will be the dynamic properties of the more mature web of 2015? Will it be like a large collective organism with a relatively static or slow-moving large-scale structure, but with a large internal information flow, and enormous activity on small timescales? Will automated tools permit construction of an annotated subnetwork of components and relationships for an arbitrary topic?

3 arXiv BACKGROUND AND LESSONS

The arXiv (arXix.org, n.d.) is an automated repository of roughly 250,000 full text research articles (as of mid-2003) in physics and related disciplines, going back over a decade and growing at a rate of 40,000 new submissions per year. (For some general background, see Ginsparg (1996a).) It began in 1991 as an e-mail interface to create, maintain, and access a set of documents for specialists in a particular subject area. Mirroring the internet growth pattern of the past 13 years described in the previous section, it added a web interface in 1993, and its usage and breadth of coverage expanded throughout the 1990's, and continues today. It now serves over 10 million requests per month (Ginsparg, 2002b), including tens of thousands of search queries per day, and over 20 million full text downloads during calendar year '02. It is a significant example of a Web-based service that has changed the practice of research in a major scientific discipline. It now provides nearly comprehensive coverage of large areas of physics, and serves as an on-line seminar system for those areas. In what follows, I will summarize some of the features of the system that depend on technologies in place by the late 1980's, and their further embellishments through the 1990's, to try to anticipate directions in scholarly publishing over the coming decade.

The arXiv operates without the editorial operations associated to peer review. As a pure dissemination system, it operates at a factor of 100 to 1000 times lower in cost than a conventionally peer-reviewed system (Ginsparg, 2001). This is the real lesson of the move to electronic formats and distribution: not that everything should somehow be free, but that with many of the production tasks automatable or off-loadable to the authors, the editorial costs will then dominate the costs of an unreviewed distribution system by many orders of magnitude. This is the subtle difference from the paper system, in which the expenses directly associated to print production and distribution were roughly the same order of magnitude as the editorial costs. When the two were comparable in cost, it wasn't as essential to ask whether the production and dissemination system should be decoupled from the intellectual authentication system. Now that the former may be feasible at a cost of less than 1% of the latter, the unavoidable question is whether the utility provided by the latter, in its naive extrapolation to electronic form, continues to justify the associated time and expense. Since many communities rely in an essential way on the structuring of the literature provided by the editorial process, a first related question is whether some hybrid methodology might provide all of the benefits of the current system, but for a cost somewhere in between the greater than $1000/article cost of current editorial methodology and the less than $10/article cost of a pure distribution system (Ginsparg, 2001). A second question is whether a hybrid methodology might also be better optimized for the differing needs, on differing timescales, of expert readers on the one hand and neophytes on the other.

When the arXiv was initiated in 1991, no physics journals were yet on-line. Its original intent was not to supplant journals, but to provide equal and uniform global access to prepublication materials (originally it was only to have had a three month retention time). Due to the multi-year period from '91 until established journals did come on-line en masse, the arXiv de facto took on a much larger role, by providing the unique on-line platform for near-term (5-10 yr) "archival" access. Electronic offerings have of course become commonplace since the early 1990's: many publishers now put new material on-line in e-first mode, and the searchability, internal reference linking, and viewable formats they provide are at least as good as those of the automated arXiv. These conventional publishers are also set up to provide superior services wherever manual oversight, at additional cost, can improve on the author's product: e.g., correcting bibliographic errors and standardizing the front- and back-matter for automated harvesting. (Some of these costs may ultimately decline or disappear, however, with a more standardized "next-generation" document format, and improved authoring tools to produce it – developments from which automated distribution systems will benefit equally.)

We can now consider the current roles of the arXiv and of the on-line physics journals and assess their overlap. Primarily, the arXiv provides instant pre-review dissemination, aggregated on a field-wide basis, a breadth far beyond the capacity of any one journal. The journals augment this with some measure of authentication of authors (they are who they claim to be), and a certain amount of quality control of the research content. This latter, as mentioned, provides at least the minimum certification of "not obviously incorrect, not obviously uninteresting"; and in many cases provides more than that, e.g., those journals known to have higher selectivity convey an additional measure of short-term prestige. Both the arXiv and the journals provide access to past materials; and one could argue that arXiv benefits in this regard from the post facto certification functions provided by the journals. It is occasionally conjectured that organized journals may be able to provide a greater degree of long-term archival stability, both in aggregate and for individual items, though looking a century or more into the future this is really difficult to project one way or another.

With conventional overlapping journals having made so much on-line progress, does there remain a continued role for the arXiv, or is it on the verge of obsolescence? Informal polls of physicists suggest that it remains unthinkable to discontinue the resource, that it would simply have to be reinvented because it plays some essential role not fulfilled by any other. Hard statistics substantiate this: over 20 million full text downloads during calendar year ’02, on average the full text of each submission downloaded over 300 times in the 7 years from ’96-’02, and some downloaded in the tens of thousands of times. The usage is significantly higher than comparable on-line journals in the field, and, most importantly, the access numbers have accelerated upwards as the conventional journals have come on-line over the past seven years. This is not to suggest, however, that physicist users are in favor of rapid discontinuation of the conventional journal system either.

What then is so essential about the arXiv to its users? The immediate answer is “Well, it’s obvious. It gives instant communication, without having to wait a few months for the peer review process.” Does that mean that one should then remove items after some fixed time period? The answer is still “No, it remains incredibly useful as a comprehensive archival aggregator,” i.e., a place where for certain fields instead of reading any particular journal, or set of journals, one can browse or search and be certain that the relevant article is there, and if it’s not there it’s because it doesn’t exist.

It has been remarked (Brinkman, 2002) that physicists use the arXiv site and do not appear concerned that the papers on it are not refereed. The vast majority of submissions are nonetheless submitted in parallel to conventional journals (at no “cost” to the author), and those that aren’t are most frequently items such as theses or contributions to conference proceedings that nonetheless have undergone some effective form of review. Moreover, the site has never been a random UseNet newsgroup-like free-for-all. From the outset, a variety of heuristic screening mechanisms have been in place 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 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. Virtually all are in that grey area of decidability, and virtually none are entirely useless to active physicists. That is probably why expert arXiv readers are eager and willing to navigate the raw deposited material, and greatly value the accelerated availability over the filtering and refinement provided by the journal editorial processes (even as little as a few months later).

4 NEW SCHOLARLY PUBLICATION MODELS

The question for our scholarly research communications infrastructure is: if we were not burdened with the legacy print system and associated methodology, what system would we design for our scholarly communications infrastructure? Do the technological advances of the past decade suggest a new methodology that provides greater utility to the research enterprise at the same or lower cost?

4.1 Open Access

There has been much recent discussion of free access to the on-line scholarly literature. It is argued that this material becomes that much more valuable when freely accessible (Berry, 2001), and moreover that it is in public policy interests to make the results of publicly funded research freely available as a public good (Bachrach, Berry, Blume, von Foerster, Fowler, Ginsparg, et al., 1998). It is also suggested that this could ultimately lead to a more cost-efficient scholarly publication system. The response of the publishing community has been that their editorial processes provide an essential service to the research community, that these are labor-intensive and hence costly, and that even if delayed, free access could impair their ability to support these operations. (Or, in the case of commercial publishers, reduce revenues to below the profit level necessary to satisfy their shareholders or investors.) Informal surveys (e.g., Ginsparg, 2001) of medium- to large-scale publishing operations suggest a wide range in revenues per article published, from the order of $1000/article to more than $10,000/article. The smaller numbers typically come from non-profit operations that provide a roughly equivalent level of service, and hence are more likely representative of actual cost associated to peer reviewed publication. Even some of these latter operations are more costly than might ultimately be necessary, due to the continued need to support legacy print distribution, but the savings from eliminating print and going to an all-electronic in-house work-flow are estimated for a large non-profit publisher to be at most on the order of 30%. (This estimate is for the American Physical Society, which publishes over 14,000 articles per year, and derives from figures discussed with its publications oversight committee. The percentage estimated for other publishing operations will vary, especially when editorial time and overhead is differentially accounted.

In the discussion that follows, however, it matters only that there will be no *windfall* savings to publishers from going all-electronic, while employing the same overall labor-intensive methodology.) The majority of the expenses are for the non-automatable editorial oversight and production staff: labor expenses that are not only unaffected by the new technology but that also increase faster than the overall inflation rate in developed countries.

It is also useful to bear in mind that much of the current entrenched methodology is largely a post World War II construct, including both the largescale entry of commercial publishers and the widespread use of peer review for mass production quality control. It is estimated that there are well over $8 billion/year in revenues in STM (Scientific, Technical, and Medical) primary publishing, for somewhere on the order of 1.5-2 million articles published/year. If non-profit operations had the capacity to handle the entirety, and if they could continue to operate in the $500-$1500 revenue per published article range, then with no other change in methodology there might be an immediate 75% savings in the system, releasing well over $5 billion globally.

One proposal to continue funding the current peer-review editorial system is to move entirely from the subscription model to an "author-subsidy" model, in which authors or their institutions pay for the material, either when submitted or when accepted for publication, and the material is then made freely available to readers. While such a system may prove workable in the long-run, it is difficult to impress upon authors the near-term advantages of moving in that direction. It would have the very useful effect of making more manifest directly to authors not only what the minimum real costs are, but also what are the hierarchies of cost within the system. This could help to bring market forces to bear on a system that currently operates on a monopolistic basis. A few examples of recently created journals experimenting with this mode are the New Journal of Physics (New Journal of Physics, n.d.), the BioMed Central journals (BioMed Central, n.d.), and the Public Library of Science journals (Public Library of Science, n.d.).

From the institutional standpoint, it would also mean that institutions that produce a disproportionate amount of quality research would pay a greater *percentage* of the costs. Some could consider this unfair, though in the long-term a fully reformed and less expensive scholarly publication system should nonetheless offer real savings to those institutions, since they already carry the highest costs in the subscription model. Another short-term difficulty with implementing such a system is the global nature of the research enterprise, in which special dispensation might be needed to accommodate researchers in developing countries, operating on lower funding scales. Correcting this problem could entail some form of progressive charging scheme and a proportionate increase in the charges to authors in developed countries, increasing the psychological barrier to moving towards an author-subsidy system. A system in which editorial costs are truly compensated equitably would also involve a charge for manuscripts that are rejected (sometimes these require even more editorial time than those accepted), but implementing that is also logistically problematic.

4.2 Peer Review

Many participants in the current peer review system regard it as the only possible quality control mechanism for the literature, signalling important contributions to readers, and necessary for deciding job and grant allocations. But this viewpoint relies on two very strong implicit assumptions: a) that the necessary signal results directly from the peer review process itself, and b) that the signal in question could *only* result from this process. The question is not whether we still need to facilitate *some* form of quality control on the literature; it is instead whether given the emergence of new technology and dissemination methods in the past decade, is the current implementation of peer review still the most effective and efficient means to provide the desired signal?

Appearance in the peer-reviewed journal literature certainly does not provide sufficient signal: otherwise there would be no need to supplement the publication record with detailed letters of recommendation and other measures of importance and influence. On the other hand, the detailed letters and citation analyses *would* be sufficient for the above purposes, even if applied to a literature that had not undergone that systematic first editorial pass through a peer review system. This exposes one of the hidden assumptions in the above: namely that peer-reviewed publication is a prerequisite to entry into a system that supports archival availability and other functions such as citation analysis. In the electronic world, that is no longer necessarily the case.

The idea of using prior electronic distribution to augment the referee process goes back at least to Rogers & Hurt (1989). Proposals along the lines of decoupling peer review from arXiv distribution can be found in Ginsparg (1994), and the notion of "overlay" journals is further discussed in Ginsparg (1996a) and

Ginsparg (1996b). A review of various “decoupling” and “author subsidy” models proposed in the mid to late 1990’s, taking advantage of new technology to implement improvements in research communication, can be found in Gass (2001). In particular, the “eprint moderator model” (Stern, 1999) was intended to reduce costs by reducing the amount of material distributed in a commercial manner.

My own experience as a reader, author, and referee in Physics suggests that current peer review methodology in this field strives to fulfill roles for two different timescales: to provide a guide to expert readers (those well-versed in the discipline) in the short-term, and to provide a certification imprimatur for the long-term. The attempt to perform both functions in one step necessarily falls short on both timescales: too slow for the former, and not stringent enough for the latter.

The observed behavior of expert readers indicates that they don’t value the extra level of filtering provided by the current review process above their preference for instant availability of material “of refereeable quality.” Non-expert readers typically don’t need the availability on the timescale of a few months, but do eventually need a much higher level of selective filtering than is provided on the short timescale. Expert readers as well could benefit on a longer timescale (say a year or longer) from more stringent selection criteria, for the simple reason that the literature of the past decade is always much larger than the “instantaneous” literature. More stringent criteria on a longer timescale would also aid significantly in the job and grant evaluation functions, for which signal on the year or more timescale remains sufficiently timely. More stringent evaluation could potentially play a far greater role than peer-reviewed publication currently does, as compared to external letters and citation analyses.

The simplest modification proposal is thus a two-tier system (for more details, see, e.g., Ginsparg (2002b)), in which on a first pass only some cursory examination or other pro forma certification is given for acceptance into a standard tier. This could be minimally labor-intensive, perhaps relying primarily on an automated check of author institutional affiliation, prior publication record, research grant status, or other related background; and involve human labor primarily to adjudicate incomplete or ambiguous results of an automated pass. The standard tier availability could also be used to collect confidential commentary from interested readers so that eventual referees would have access to a wealth of currently inaccessible information held by the community, and help to avoid duplication of effort. Then at some later point (which could vary from article to article, perhaps with no time limit), a much smaller set of articles would be selected for the full peer review process. The initial selection criteria for this smaller set could be any of a variety of impact measures, to be determined, and based explicitly on their prior widespread and systematic availability and citability: e.g., reader nomination or rating, citation impact, usage statistics, editorial selection, … .

The precise criteria would depend on the architectural details of the repositories. In a federation of institutionally and disciplinarily held repositories, the institutional repositories (e.g., Dspace (Dspace Federation, n.d.)) could rely on some form of internal endorsement, while the disciplinary aggregates could rely either on affiliation or on prior established credentials (termed “career review” in Kling, Spector, & McKim (2002), as opposed to “peer review”). Alternate entry paths for new participants, such as referrals from prior credentialed participants or direct appeal for cursory editorial evaluation (not full-fledged peer review), would also be possible. The essential idea is to facilitate communication within the recognized research community, without excessive noise from the exterior (Ginsparg, 1994). While multiple logically independent (though potentially overlapping (Ginsparg, 2001)) upper tiers could naturally evolve, only a single globally held standard tier is strictly necessary, with of course any necessary redundancy for full archival stability. Suitable licensing procedures or copyright retention (Bachrach et al., 1998) to facilitate such a system are consistent with the spirit of copyright law, “To promote the Progress of Science and useful Arts” (for a recent discussion, see Willinsky (2002)).

Recent experience in physics and related disciplines continues to reinforce the desirability of experimentation within this model space, with the expectation that similar implementations will prove feasible in other disciplines.

5 CONCLUSION

Just as some of the trends of the early 21st century could have been anticipated in the late 1980’s, the networked world in 2015 should seem familiar in many ways . But we’re likely to be surprised both by qualitatively new network applications, and by “threshhold effects” (or “tipping points”) at which a certain level of community participation, or of relevant data or metadata available, or of computing power available, fundamentally changes the way we view or interact with a set of resources.

A few incipient examples of the latter include resources such as the American Physical Society's "PROLA" (prola.aps.org), containing the scanned and OCR'd full text of all of its journals back to their inception in 1893, the JSTOR project (www.jstor.org), similarly comprising back issues of mainly non-science journals, and the Astrophysical Data System (wwwads.harvard.edu), 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. Continued development of navigational tools — such as improved document clustering algorithms, e.g., for scholarly purposes those of Vivisimo (Vivisimo, n.d.) — will continue to scale up expectations for how quickly and easily an increasing breadth and depth of information and data can be accessed and absorbed.

But we're still just scratching the surface of what can be done with large and comprehensive full text aggregations. A forward-looking example is the PubMed Central database (PubMed Central, n.d.), operated in conjunction with GenBank and other biological databases at the U.S. 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. Proteins 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 of the relevant standard biology or biochemistry textbook in the books database. We see that the enormously powerful sorts of datamining 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. Ongoing experiments with the arXiv database (described in section 3) include citation prediction tasks, usage prediction and evaluation, datacleaning, subject area structuring, and alternate evaluations of significance (see, e.g., KDD Cup (2003)). These applications and others all depend on the nascent presence of openly available and comprehensive databases with large dedicated communities of users and contributors.

With respect to scholarly publication, there is something of an ironic ending to the speculations here concerning the future. The modifications sketched in section 4 are intended as a starting point for discussion of how recent technological advances could be used to improve the implementation of peer review. They are not intended to be revolutionary, but sometimes a small adjustment, with seemingly limited conceptual content, can have an enormous effect. We recall that commercial publishers entered the realm of scholarly publishing to fulfill an essential role during the post World War II period, precisely because the non-profits did not have the requisite capacity to handle the dramatic increase in STM publishing with then-available technology. An altered methodology based on the electronic communications networks that evolved through the 1990's could prove better scalable to larger capacity. In this case, the technology of the 21st century would allow the traditional players from a century ago, namely the professional societies and institutional libraries, to return to their dominant role in support of the research enterprise.

In that case, some things in 2015 could be closer to the way they were a century ago than they were in 1989!

6 REFERENCES

Arms, W. (2000) How Effectively Can Computers Be Used for the Skilled Tasks of Professional Librarianship? *D-Lib Magazine 6*(7/8). Retrieved December 1, 2003 from:
http://www.dlib.org/dlib/july00/arms/07arms.html

arXiv.org (n.d.) Homepage of arXiv.org e-Print archive. Available from: http://arXiv.org/

Bachrach, S., Berry, R.S., Blume, M., von Foerster, T., Fowler, A., Ginsparg, P., Heller, S., Kestner, N, Odlyzko, A., Okerson, A., Wigington, R., Moffat, A. (1998) Who should own scientific papers? *Science 281*, 1459–1460. Retrieved December 1, 2003 from:
http://www.sciencemag.org/cgi/content/full/281/5382/1459

Berry, R. S., Alexander, S. A., Allen, B. E., Baumgardner, M., Branscomb, A. W., Cizewski, J. A., Dubetz, M. W., Faulhaber, G. R., Gabrynowicz, J. I., Ginsparg, P., Gordon, W. E., Hallgren, R. E., King, D. W., Krichevsky, M. I., Malone, T. F., Melillo, J. M., Reichman, J. H., Richard, B. K., Schreier, E. J., Söll, D., Westbrook, J. H., Wigington, R. L., Uhlir, P. F. (1997) Bits of power: issues in global access to scientific data (Committee on issues in the transborder flow of Scientific Data, U.S. National Committee for CODATA, Commission on Physical Sciences, Mathematics, and Applications, National Research Council). Retrieved December 1, 2003 from the National Academies Press website: http://www.nap.edu/readingroom/books/BitsOfPower/

Berry, R.S. (2001) Is electronic publishing being used in the best interests of science? The scientist's view. In Elliot, R. & Shaw, D. (Eds.) *Proceedings of the second ICSU Press-UNESCO Conference on Electronic Publishing in Science*. Retrieved December 1, 2003 from: http://users.ox.ac.uk/ icsuinfo/berryfin.htm

BioMed Central (n.d.) Homepage of BioMed Central. Available from: http://www.biomedcentral.com

Brinkman, W. (2002, January) Brinkman Outlines Priorities, Challenges for APS in 2002, *APS News Online*. Retrieved December 1, 2002 from WorldWideWeb: http://www.aps.org/apsnews/0102/010208.html

Dspace Federation (n.d.) Homepage of Dspace. Available from: http://www.dspace.org/

Gass, S. (2001) Transforming Scientific Communication for the 21st Century. *Science and Technology Libraries, 19*(3/4) 3–18.

Ginsparg, P. (1994) First Steps Towards Electronic Research Communication. *Computers in Physics, 8*(4), 390 (see also P. Ginsparg, After Dinner Remarks, http://arXiv.org/blurb/pg14Oct94.html, presented at the APS e-print Workshop at Los Alamos National Laboratory. Retrieved December 1, 2003 from: http://publish.aps.org/EPRINT/KATHD/toc.html

Ginsparg, P. (1996a) Winners and losers in the global research village. In Elliot, R. & Shaw, D. (Eds.) *Proceedings of the ICSU Press-UNESCO Conference on Electronic Publishing in Science*. Retrieved December 1, 2003 from: http://users.ox.ac.uk/ icsuinfo/ginsparg.htm
(copy at http://arXiv.org/blurb/pg96unesco.html)

Ginsparg, P. (1996b, November) Los Alamos XXX. *APS News Online*. Retrieved December 1, 2003 from: http://www.aps.org/apsnews/1196/11718.html (copy at http://arXiv.org/blurb/sep96news.html)

Ginsparg, P. (2001) Creating a Global Knowledge Network. In Elliot, R. & Shaw, D. (Eds.) *Proceedings of the second ICSU Press-UNESCO Conference on Electronic Publishing in Science*. Retrieved December 1, 2003 from: http://users.ox.ac.uk/ icsuinfo/ginspargfin.htm
(copy at http://arXiv.org/blurb/pg01unesco.html)

Ginsparg, P. (2002a) Scholarly Information Architecture. *CODATA 2002: Frontiers of Scientific and Technical Data*. Montreal, Canada. Retrieved December 1, 2003 from the *CODATA* website:
http://www.codata.org/codata02/codata2015/index.html

Ginsparg, P. (2002b) Can Peer Review be better Focused? *Science & Technology Libraries 22*(3/4) 5–18 (copy at http://arXiv.org/blurb/pg02pr.html)

KDD Cup (2003) Homepage of KDD Cup 2003. Retrieved December 1, 2003 from the *Department of Computer Science, Cornell University*, website: http://www.cs.cornell.edu/projects/kddcup/

Kling, R., Spector, L., & McKim, G. (2002, August) Locally Controlled Scholarly Publishing via the Internet: The Guild Model, *The Journal of Electronic Publishing*. Retrieved December 1, 2003 from: http://www.press.umich.edu/jep/08-01/kling.html

NCBI DTD (n.d) Homepage of Archiving and Interchange DTD. Retrieved December 1, 2003 from the *National Library of Medicine* website: http://dtd.nlm.nih.gov/

New Journal of Physics (n.d.) Homepage of the New Journal of Physics. Available from:
http://www.njp.org/

Public Library of Science (n.d.) Homepage of the Public Library of Science. Available from:
http://www.plos.org/

PubMed Central (n.d.) Homepage of PubMed Central. Retrieved December 1, 2003 from the *National Library of Medicine* website: http://www.pubmedcentral.nih.gov/

Rogers, S. & Hurt, C. (1989, October 18) How Scholarly Communication Should Work in the 21st Century, *The Chronicle of Higher Education*, A56.

Stern, D. (1999) eprint Moderator Model. Retrieved December 1, 2003 from *Yale Science Libraries* website: http://www.library.yale.edu/scilib/modmodexplain.html (version dated Jan 25, 1999)

Vivisimo (n.d.) Homepage of Vivisimo. Available from: http://vivisimo.com/

Willinsky, J. (2002) Copyright Contradictions in Scholarly Publishing. *First Monday, 7*(11). Retrieved December 1, 2003 from: http://firstmonday.org/issues/issue7_11/willinsky/index.html