COMMENT

HARPERCOLLINS

ENVIRONMENT Classic children's eco fable turns 40 **p.148**

HISTORY The travails of the first mission to calibrate latitude **p.149**

EXHIBITION Images of the mind, from Golgi to Gormley on show in Dresden **p.151**

POLICY Defending US recommendations for lab animal use and care **p.152**

The arXiv server in the early 1990s: a computer that helped to change the world of physics.

J. FLOWER/LANL

ArXiv at 20

Paul Ginsparg, founder of the preprint server, reflects on two decades of sharing results rapidly online — and on the future of scholarly communication.

Twenty years ago this month, I launched an electronic bulletin board intended to serve a few hundred friends and colleagues working in a subfield of theoretical high-energy physics. I had recently moved to the Los Alamos National Laboratory in New Mexico and for the first time had my own computer on my desk, and the desire to simplify the exchange of unpublished manuscripts (preprints) between researchers, previously distributed as paper copies by post.

This automated repository and alert system for physics preprints, at hep-th@xxx.lanl.gov, was implemented shortly before the dawn of the web era. As I e-mailed to a colleague at CERN more than a year later: 'I know nothing of WWW, what is it?' The original plan was for roughly 100 full-text article submissions every year, each stored for three months until the existing paper distribution system could catch up. By popular demand, nothing was ever deleted.

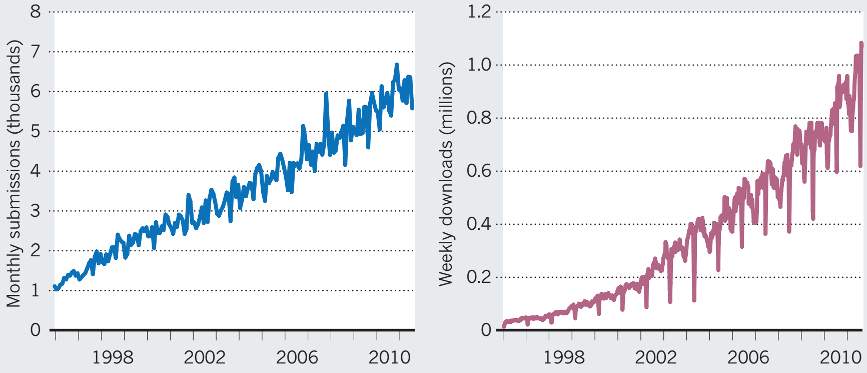
Within a few years it had evolved into a web resource at arXiv.org that now contains close to 700,000 full texts, receives 75,000 new texts each year, and serves roughly 1 million full-text downloads to about 400,000 distinct users every week (see graphs). It has broadened, first to cover most active research fields of physics, then to mathematics, nonlinear sciences, computer science, statistics and, more recently, to host parts of biology and finance infiltrated by physicists.

It is heartening, 20 years later, to see a stable and successful arXiv, running some of the original software and providing services to a community nearly a thousand times larger than expected. But at some point a thorough overhaul will be needed to keep pace with new online trends and opportunities.

For me, the repository was supposed to be a three-hour tour, not a life sentence. ArXiv was originally conceived to be fully ▶

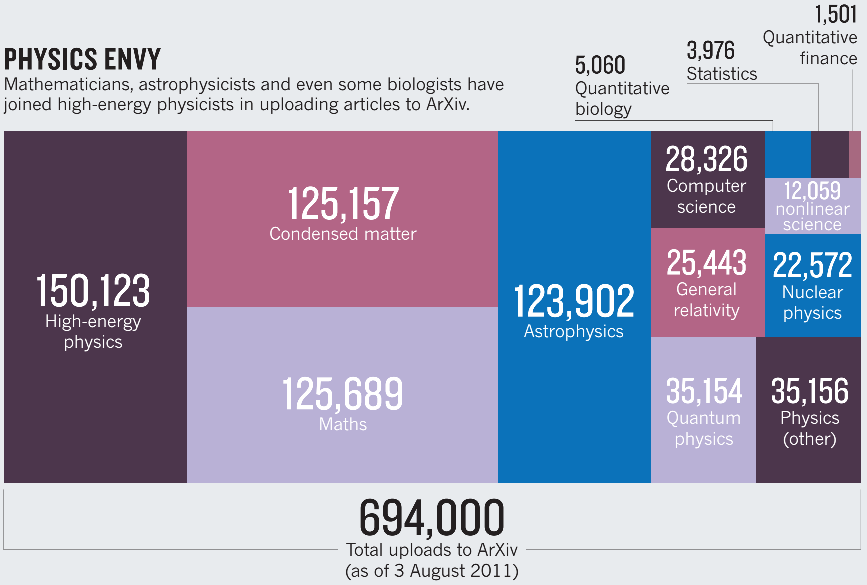
DIGITAL PIONEERS LEAD THE WAY TO SHARING RESEARCH ONLINE

The popularity of the arXiv preprint server has grown inexorably since its launch in the early 1990s. Academics enjoy the universal access, low cost and speed of online distribution.



PHYSICS ENVY

Mathematicians, astrophysicists and even some biologists have joined high-energy physicists in uploading articles to ArXiv.



SOURCE: ARXIV

automated, so as not to scuttle my research career. But daily administrative activities associated with running it can consume hours of every weekday, year-round without holiday. So, from September, the site will be entirely in the hands of the staff of Cornell University Library in Ithaca, New York. I will remain on the advisory board and continue some research projects in text and data mining and in supporting next-generation document formats and information filters.

To reflect on ArXiv is to reflect on a research world transformed by a two-decade revolution in information technology, with vast quantities of literature and associated resources now available on demand. Yet, it is a surprise that scholarly publishing as a whole remains in transition. There is no consensus on the best way to implement quality control (top-down or crowd-sourced, or at what stage), how to fund it or how to integrate data and other tools needed for scientific reproducibility.

My hope is that rather than merely using electronic infrastructure as a more efficient means of distribution, the revolution-in-waiting will ultimately lead to a more powerful knowledge structure, fundamentally transforming the ways in which we process and organize scientific data.

DEMOCRATIC GOALS

The original bulletin board was engineered to level the research playing field. It is hard to imagine now, but considerable time and effort was once spent printing, photocopying and posting preprints to a privileged few friends and colleagues, before publication in formal journals. The idea of a central repository was to allow any researcher worldwide with network access to submit and read full-text articles, giving equal entry to everyone from graduate students up. (The early Internet was an academic playground — the general public didn't start coming online until a few years later.)

Within two years, arXiv had evolved into the primary daily resource for a global community of researchers. It became a place to stake intellectual precedence claims, catalysing further growth.

Launched in 1991, before any conventional journals were online, arXiv pioneered many of the tools now taken for granted. We led the way in using the abstract page as a hub to diverse formats and resources, linking author names automatically to search functions, and we were early adopters of electronic formats from compressed postscript to PDF for file sharing. The repository showed that researchers were willing and eager to move to fully electronic means of dissemination. It also foreshadowed the 'interactive web', insofar as it provided a rudimentary framework for users to deposit content. The value of this content is then amplified by sharing.

In many ways, building the technology was simpler than managing the sociological and financial aspects. A decade ago, the main site moved with me and became embedded within Cornell University library. Although it serves more users on a daily basis than any other library resource, most of those users are external, so it is less clear where it should fit in university funding priorities. In the absence of a wealthy donor willing to provide a small endowment in exchange for far more name recognition than any traditional building donation (hint, hint), the library has recently asked institutions who are heavy users to contribute to running costs (http://arxiv.org/help/support). This distributes the financial burden and oversight to a larger community, while buying time to investigate long-term business models.

Physicists were quick to adopt widespread sharing of electronic preprints, but other researchers remain reluctant to do so. Fields vary widely in their attitudes to data and ideas before formal review, and in choosing to share electronic preprints, each community will have to develop policies and protocols best suited to their users. A talk I gave in 1997 to a group of biologists helped catalyse the resource now known as PubMedCentral — run by the US National Institutes of Health. I served on the initial advisory board, which soon decided not to host any unrefereed materials, even carefully quarantined, in part for fear of losing essential publisher participation. There remain many legitimate reasons for individual researchers to prefer to delay dissemination, from uncertainty over correctness, to retaining extra time for follow-ups, to sociological differences in the way publication is regarded — in certain fields, the research somehow doesn't count until peer reviewed.

No community that has adopted arXiv usage has renounced it, however, so the growth has been inexorable. Adoption by some fields, including computer science, started as a trickle, increasing dramatically many years later. Even some subfields of physics have experienced delayed adoption: emerging research in 2008 into superconductivity in iron compounds brought in a group of condensed-matter experimentalists traditionally more cautious about disseminating results. They were ultimately won over by the need to stake precedence claims and get their results in front of theorists.

Even today, fields vary hugely in how they recognize intellectual precedence. It baffles me that scientists in some fields can announce a result in a public forum, such as a meeting, while another group can reproduce the results, publish first in a journal, and be given complete intellectual precedence, as though the information did not exist until vetted by the referee process. Journal editors and referees should make more effort to ensure proper attribution is given to publicly accessible materials in a stable resource, such as arXiv.

WHERE NEXT?

As the arXiv community has diversified, so have its desires. Some users have requested support for comment threads related directly to papers on the site, while others prefer that it maintain its unadulterated stream of author-provided content. I have sympathy for more interactivity: in today's social web, a one-way channel seems an anachronism. But because maintaining utility and civility in online discussions can be labour-intensive, our policy has been that such services should be external to the main repository. The same considerations apply to self-organized, or 'crowd-sourced' forms of review. In recent years, some external blogs have begun to host useful comment threads, linked back from arXiv abstract pages through a trackback mechanism, but I don't predict that dedicated blogging by individual scientists will grow substantially. More scalable tools for discussion are provided by question-and-answer sites such as mathoverflow.net, where expert mathematicians, in the course of answering one another's posted questions, provide links to maths articles hosted on arXiv.

Again, because of cost and labour overheads, arXiv would not be able to implement conventional peer review. Even the minimal filtering of incoming preprints to maintain basic quality control involves significant daily administrative activity. Incoming abstracts are given a cursory glance by volunteer external moderators for appropriateness to their subject areas; and various automated filters, including a text classifier, flag problem submissions. Although the overall rate of such submissions is well below 1%, they tend to cluster in specific areas (such as general relativity, quantum mechanics and unified theories in physics; proofs of the Riemann hypothesis, Goldbach's conjecture and new proofs of Fermat's last theorem in mathematics; P versus NP problem in computer science).

➲ NATURE.COM
Discuss this article online at:
go.nature.com/m3u6oj

Moderators, tasked with determining what is of potential interest to their communities, are sometimes forced to ascertain 'what is science?' At this point arXiv unintentionally becomes an accrediting agency for researchers, much as the Science Citation Index became an accrediting agency for journals, by formulating criteria for their inclusion. Although decisions are biased towards permissiveness, inevitably some authors object that it is never permissive enough.

DIGITAL GENERATION

The idea that print journals had outlived their usefulness was already in the air in the early 1990s. David Mermin memorably wrote in *Physics Today* in 1991: "The time is overdue to abolish journals and reorganize the way we do business."[1] By the mid 1990s, it seemed unthinkable that free and unfettered access to non-refereed papers on arXiv would continue to coexist indefinitely with quality-controlled but subscription-based publications. Fifteen years on, researchers continue to access both, successfully compartmentalizing their different roles in scholarly communication and reward structures.

The transition to article formats and features better suited to modern technology than to print on paper has also been surprisingly slow. Page markup formats, such as PDF, have only grudgingly given way to XML-based ones that support features such as manipulable graphics, dynamic views, linked annotations and semantic markup. Part of this caution is a result of the understandable need to maintain a stable archive of research literature, as provided by paper over centuries.

> ***"It is a surprise that scholarly publishing as a whole remains in transition."***

Configuring scholarly communication infrastructure for the next generation of researchers requires getting into the heads of current undergraduates and graduate students. Their life experience is of immediate online availability and global search engines, and they arrive imbued with the social-network mentality of sharing links, photos, videos and status updates. Yet, my own informal survey of graduate students reveals information-gathering techniques familiar to most older scientists. Students still follow citation trees, search by keywords and consult with peers and mentors, with the latter as important as ever for weeding out unreliable sources. Students also say that they search preferentially for open-access resources when working from home, because accessing subscription-based journals, even when available through an institutional proxy, can be frustratingly painful.

Navigating increasing quantities of data inevitably raises concerns of information overload. This phenomenon is documented back to the dawn of writing, accelerated with the invention of the printing press, and has been re-emphasized by every generation since. The superficial response is a call for better filters, but an imperfect filter can be more harmful than none. For example, commonly used recommender systems based on passive measures of global popularity can broaden individual reading choices, but effectively broaden everyone in the same direction, thereby leading to less overall community diversity. (The cynic would say reinforcing faddishness in already faddish fields.)

On arXiv, we have seen some of the unintended effects of an entire global research community ingesting the same information from the same interface on a daily basis. The order in which new preprint submissions are displayed in the daily alert, if only for a single day, strongly affects the readership on that day and leaves a measurable trace in the citation record fully six years later[2,3]. Some researchers, wise to this, time their submissions to arrive just after the daily afternoon deadline to maximize their prominence in the next day's mailing. Filters that highlighted 'popular' materials over longer periods of time would exacerbate this effect. Hence any recommender system on arXiv would need, at minimum, to be personalized to individual readership preferences and interests to reduce herding behaviour. Experiments with such systems are ongoing, and may be put online within a year or two if they perform properly.

For now, the open questions of arXiv's long-term role and its relationship to conventional publishing, the details of its funding model, and its overall intellectual supervision, are to be resolved in coordination with its users and stakeholders. A meeting of international sponsoring institutions will be hosted by the Cornell Library next month to discuss the transition of arXiv to a collaboratively governed, community-supported resource. It will be a challenge to keep it attuned to the needs of future generations of researchers.

My hope is that the barrier to implementation of new ideas in this realm will remain low enough that, if all else fails, some young researcher elsewhere can launch another tiny ship on a fateful trip. ■

Paul Ginsparg *is at the Department of Physics, Cornell University, Ithaca, New York 14853, USA.*
e-mail: ginsparg@cornell.edu

1. Mermin, N. D. *Phys. Today* **44,** 9 (1991).
2. Haque, A. & Ginsparg, P. *J. Am. Soc. Inf. Sci. Technol.* **60,** 2203–2218 (2009).
3. Haque, A. & Ginsparg, P. *J. Am. Soc. Inf. Sci. Technol.* **61,** 2381–2388 (2010).