From the Los Alamos Preprint Archive to the arXiv: An Interview with Paul Ginsparg

Electronic dissemination of research findings has long interested science editors, but many of us in CSE know relatively little about such dissemination in fields other than biology and medicine. Therefore, for this issue of Science Editor, *I interviewed physicist* **Paul Ginsparg**, *who in 1991 developed the Los Alamos Electronic Preprint Archive, recently renamed arXiv. I appreciate his having answered my questions while busy moving from Los Alamos National Laboratory (LANL) to Cornell University, where he is continuing to maintain the arXiv in conjunction with the Cornell University Library.*

Barbara Gastel

What is the arXiv? How can one access it?

The arXiv is an automated electronic repository that permits researchers to deposit their full-text research articles, including all graphics, and permits interested parties to access them free of charge. The articles are typically posted either before, during, or after peer review, at author discretion. The arXiv can be accessed via the World Wide Web at *arXiv.org/* or by its historical e-mail and ftp interfaces at the same address. It includes a subscription list that provides a subject-based alert system for new submissions.

How did the idea for the arXiv arise? How quickly did it catch on?

The idea germinated for a few years. By the middle 1980s, most physicists and mathematicians were using the scientific typesetting language TeX to produce their documents, and we'd switched from using the telephone to e-mail for much of our communication.

In 1987, I'd mentioned to librarian Louise Addis at the Stanford Linear Accelerator Laboratory (SLAC) library, in charge of a title-author indexing system for high-energy physics known as SLAC-Spires, that they should consider maintaining an electronic repository of the full texts in TeX. She said it was something they'd love to do but would require too much manual labor beyond what they were already doing. Neither of us thought in terms of a fully automated system; moreover, as a faculty member at Harvard at the time, I wouldn't have had the elective time to pursue it anyway.

By 1991, full-text articles in my field of string theory were being regularly e-mailed to a mailing list, and at Aspen that summer a physicist commented about being inundated with these "large" files (actually much smaller than the typical .doc or .pdf attachment these days). By then I had my own workstation rather than a shared mainframe, knew how to program it, knew how minimal the diskpace and CPU requirements of such a system would be, and, having joined LANL as a research staff member in 1990, had the time to undertake such a project.

I decided it would be feasible to set up an automated e-mail repository for the full text with an alert system that sent around only the accumulated new abstracts once a day with instructions for retrieving the full articles via automated e-mail request. Later that summer (after some travel), I spent an afternoon or two implementing the software and put it online, and it caught on immediately. I was originally anticipating about 100 submissions per year from the roughly 200 people in the one little subfield it originally covered, but there were multiple submissions per day from day 1, and by the end of the year a few thousand people were involved.

See *arXiv.org/show_monthly_submissions* for how this developed. We received 33,159 new submissions in calendar year 2001.

How does the arXiv work? For example, how are papers submitted and disseminated? How is the arXiv funded?

Articles can be submitted by e-mail, by anonymous ftp, or by Web upload. Any package of files can be submitted, and they arrive with metadata (authors, title, abstract, and so on) in a specified format for use in generating the search indexes. For the last few years it had been supported by a combination of National Science Foundation (NSF), Department of Energy (DOE), and LANL library funds. At Cornell it will be supported instead by a combination of NSF and Cornell University Library funds (that is, no longer DOE funds).

How has the arXiv evolved over the years?

In 1993, as Web browsers became more commonplace, we added a Web interface to the original e-mail interface. Most of the other changes have been incremental: better autoprocessing of submissions, improved indexing and searching, addition of the international mirror network. The basic core operations and underlying philosophy have remained unchanged.

What is your role in the arXiv? What do you do to maintain it? Are others involved, and what are their roles?

Since 1993, when the NSF funding started, I've typically employed two people to help with software development and provide an e-mail "help desk" for occasional questions that arise that need personal intervention. My own technical role in the last few years has been minimal—not much time left after securing funding and giving presentations at meetings (this is sad because designing and writing software were my only real talents in any of this).

How has disseminating papers on the arXiv related to publishing papers in journals? For example, does inclusion in the arXiv tend to replace publication in a journal?

Interview with Paul Ginsparg continued

Does it tend to precede it?
As mentioned earlier, authors are free to submit either before or after journal submission (or not submit to a journal at all). Some journals permit electronic submission directly from the arXiv; that is, they have a submission form that permits simply specifying the arXiv-assigned identifier.

Many high-energy physicists have asserted that the journals are less relevant and they can get by on the arXiv alone, and there still remains much truth to that, at least as far as communication of research results goes.

But there remains much lingering conservatism in the system: If people still need the "peer-reviewed" publications for grants and jobs, and moreover if it's relatively painless, then why not take that additional small step? It doesn't "cost" anything, and it's a form of insurance policy. I recently scanned the high-energy physics hep-th and hep-ph archives for submissions entered during 1999 and found that over 70% had an entered journal reference. (The journal references for these fields are provided by SLAC-Spires instead of relying on authors, so they're fairly well covered.) The remaining percentage includes a substantial fraction of conference proceedings and theses, so only a relatively small number were never submitted to journals or rejected by them.

My intuition is that even if the journal system were to be abandoned by this most "radical" community, some form of review system would be reinvented anyway, so it makes sense to remain in coordination with the professional societies (like the American Physical Society) so they can adiabatically evolve to where they need to go instead of having to rise from the ashes later.

Are papers edited or reviewed in any way before they appear in the arXiv? Does inclusion in the arXiv itself serve an editorial role?
No explicit editorial role; submissions are as entered by the authors. The only screening is of the e-mail address of the submitter to ensure a recognized institutional affiliation.

If you haven't yet said so: How large is the arXiv? What are the main fields represented? How prominent is the arXiv in those fields?
The ArXiv had roughly 185,000 total submissions by the end of calendar year 2001. The main fields represented are physics, mathematics, nonlinear dynamics, and computer science. The arXiv's greatest prominence is in physics, in which some subfields (such as high-energy physics, where it started) have had virtually 100% participation since the middle 1990s. The fastest growing fields since then have been astrophysics and condensed-matter physics.

Do you foresee expanding the arXiv to include fields not yet represented? If so, what might be some of the issues in doing so?
Expansion is certainly likely, and there have been many requests from representatives of fields that would like to be included. The likely problems are more sociologic than technical. For example, it seems that in the biomedical and life sciences, researchers have ceded a great deal of power to high-visibility journals that might act to suppress this alternative mode of research communication. (Certainly the "Public Library of Science" movement, *publiclibraryofscience.org*, is an alternative means of reforming some of these practices from within.)

What other changes do you foresee for the arXiv? What effect do you think your move to Cornell is likely to have?
The main effect of moving to Cornell is that the system will at last have a solid long-term institutional base. In principle, that will ultimately permit me to return full-time to my primary avocation, for which I'm somewhat better trained: physics research. The move to Cornell also enhances the possibility of expanding into other fields, since it is such a broad-based academic institution.

For those of us interested in word origins: How did you decide on arXiv?
The main site at Los Alamos had been named *xxx.lanl.gov*. This was back in 1991, long before "xxx" had acquired its current "adult" implications on the Internet. (There's also a little story behind the "xxx", but fortunately you only asked about arXiv.)

In late 1998 I decided we needed to register an ".org" domain name to facilitate rapid redirection of accesses to a different main site in the event of hardware or network problems and to be able to normalize the mirror-site names (for example, *uk.arXiv.org* for the UK site, *fr.arXiv.org* for the French site, and so on). All the *archive.org*, *archives.org*, *thearchive.org*, and *thearchives.org* domains were already taken, so I had to dream up something else. While driving up to Taos for a holiday dinner, I decided that since the word had a Greek root I could use X to indicate the Greek chi, imitating Donald Knuth's usage in the scientific typesetting language TeX (pronounced Tech). I liked being able to preserve at least one of the original three x's, and I recognized the virtues of a unique "brand name". At dinner, I wrote down "arXive" on the back of a receipt to get my wife's opinion, and she suggested eliminating the final e (as in the German archiv), and that's what I went ahead and registered a couple days later.

It looked odd at first, but people get used to these things.

Also, the word *archive* itself goes back to the Greek *archos* for ruler, or *arche* to begin or rule, where *archeion* was the government house and led to the Latin *archivum* as a place where public records or historical documents were preserved. Hence, it was natural to have started this in the .gov domain.

For those who wish to learn more about the arXiv, what sources of further information would you recommend?
My most recent writeup for a UNESCO conference on "electronic publishing in science" can be found at *arXiv.org/blurb/pg01unesco.html*, and it contains references to earlier resources.

Many thanks.