"Project Euclid and the ArXiv: Complimentary and Contrasting Elements for Sustainability,"

H. Thomas Hickerson
Cornell University

An edited version of remarks presented at the "Workshop on Sustainable Models for University-Based Scholarly Publishing," conducted at Columbia University on June 1, 2004.

Sustainability is the principal topic of today's deliberations. New models are evolving for the development and dissemination of scholarly information, but viable options are dependent on stable organizational foundations and sound managerial and financial models. My remarks today are not designed to provide an overarching vision for the future of scholarly publishing, but are intended to elucidate factors critical to the success of such visions. I am pleased to have this opportunity to review with you strategies presently being employed by the electronic publishing program of the Cornell University Library. I will focus on two particular research publishing endeavors, Euclid and the physics, mathematics, and computer science e-print arXiv. These two alternative publishing instances offer us a lens through which to analyze elements critical to sustainability. Their very different operational models illustrate differences in sustainability strategies, and yet there are important similarities between the two.

I will describe each of these initiatives briefly, but first I would like to draw your attention to points raised in a recent article that I will employ in analyzing sustainability. The article, "Building on Success, Forging New Ground: The Question of Sustainability," is by Donald Waters, Program Officer of Scholarly Communications at The Andrew W. Mellon Foundation. Appearing in the May 2004 edition of *firstmonday,* a peer-reviewed journal on the Internet, the article provides a succinct statement of Water's thoughts on sustainability of digital scholarly resources, a principal concentration of his and of the Mellon Foundation. <http://www.firstmonday.org/issues/issue9_5/waters/index.html> He identifies three principal factors:

(1) Development of such resources depends on a clear definition of the audience and the needs of users;

(2) The resource must be designed to take advantage of economies of scale;

(3) In creating an enduring resource, careful attention is needed to the design of the organization that will manage the resource over time.

Water's ideas are very similar to my own thinking on these issues. Although we might differ regarding ideal implementations, I think that the principles employed

in my review echo those espoused by Waters, and in describing the development, operation, and current status of Euclid and the arXiv, I will employ these factors in analyzing and comparing the nature and potential sustainability of each.

Euclid

Euclid is a new, emerging publishing venture. Development began in the summer of 2000, and it entered production in January of 2003. Preliminary planning was initiated in 1999 after Steven Rockey (Director of Cornell's Mathematical Library) conducted a comprehensive analysis of journal publishing in mathematics. At that time, *Mathematical Reviews,* the major U.S. indexer of mathematical literature, provided cover-to-cover indexing of 544 titles considered high-density mathematics journals. Seventeen large publishers, including Springer-Verlag, Reed Elsevier, Academic Press, the American Mathematical Society, and the Society of Industrial and Applied Mathematics, accounted for 179 titles. The remaining 365 titles were from publishers who produced three or fewer titles, the majority only producing one. Very few of these independent publishers are commercial concerns; most are scholarly societies or are associated with universities, frequently with math departments. Research in mathematics is conducted at a wide variety of institutions, and many reputable local journals are considered core publications.

Shortly after this study was conducted, a wave of commercial publishing mergers began that reduced the number of large math publishers further, and in an era of commercial consolidation, independent journals are, more than ever, important for their role in maintaining the diversity of expression that mathematicians value greatly. These journals are also valuable for their significant quality and low prices, in many cases providing serious competition for much more expensive commercial titles. Therefore, maintaining their competitive edge and economic well-being is critical to this scholarly community.

Not surprisingly, these independent publishers have been slow to move to online publishing. Many have not made the transition at all; others have moved online on a "shoestring" without an institutional commitment sufficient to address important user issues, such as searching and reference linking, or operational issues such metadata, authentication, electronic commerce, system development, and preservation. And while few journals have gone online, mathematicians have largely declined to place their writing online via other means. Although the arXiv included mathematics as one of its principal categories from the time of its founding in 1991, currently the arXiv includes less than 7% of current math publishing output (Zsuzsa Koltay and H. Thomas Hickerson, "Project Euclid and the Role of Research Libraries in Scholarly Publishing," *Journal of Library Administration,* 35:1/2, 2001, p. 87), a very different pattern from that of high-energy physics. The publishing patterns chosen by mathematicians and their publishers clearly demonstrate the impact of

disciplinary culture on the potential success of new publication alternatives. In an effort to understand such patterns, Euclid operated in a pilot phase with six publishers of twelve titles for more than a year during system development, actively soliciting their opinions regarding the most important features for inclusion in an online publishing system. Today, an advisory board, chiefly comprised of mathematicians, provides ongoing understanding of the discipline, but the board also includes university librarians and the director of the Scholarly Publishing and Academic Resources Coalition (SPARC), representing different components of the user community.

Although receiving significant assistance from The Andrew W, Mellon Foundation during its early development, Euclid has been designed to become a self-sustaining enterprise from its inception. So, where does Euclid stand today, and what are the prospects for Euclid's balancing direct-costs and revenues by the end of fiscal 2006/2007, Cornell's goal?

As of March 2004, Euclid had contracts with seventeen publishers of thirty-three titles and included 945 issues with 11,609 articles. Indicative of the current rate of growth: there were14 publishers, 22 journals, and 6,457 articles in December 2003. This near-doubling of content in three months reflects more titles coming online and also journals adding full backfiles. Publishers include academic and professional societies, math departments, university presses, and small commercial publishers. Euclid has marketing agreements with representatives for North America, Europe, Japan, and Asia-Pacific. Contractual arrangements include Euclid Prime, in which Euclid presently sells an aggregation of 18 journals, dividing the revenue with publishers with no "out of pocket" costs to publishers; Euclid Select through which titles are sold individually; and Euclid Direct in which publishers pay for hosting and other services, but market their own subscriptions. There is also a pay per view capacity for most articles. The *Annals of Mathematics,* the world's premier math journal, is distributed open access through Euclid. There are presently 65 subscribers to Euclid Prime, principally from the U.S. and Canada, but also including subscribers in Japan, Italy, and Vietnam. A substantial number of new Asian subscribers will join in 04/05, including consortial sales in South Korea, Taiwan, Singapore, and Thailand. All of this looks relatively positive, and we are proud of our progress. We are inexpensive, both in our subscription rates and our services to publishers. Revenues for the first ten months of 03/04 were $153,000, 62% from subscriptions and 38% for services. Our costs for the same period were $347,000. Economies of scale are essential. We need to double the number of publishers and subscribers quickly, and implement that growth with minimal increase in production costs. To be able to do that, we need to enhance the efficiencies of our software, generalizing the functionality to both reduce the costs of bringing up new titles and broadening the flexibility and services supported by DPubS (Digital Publishing System), the software we developed to support Euclid. We are presently seeking external funding to extend DPubS functionality and to make it Open Source. This extension will not only enhance Euclid, but it will

make it possible to easily publish in other fields and formats using the same technical infrastructure. New initiatives of this type are already in development. It is a tough challenge but an exciting one.

arXiv

Turning to the arXiv, it reflects a very different publishing model. For a variety of reasons, its creation may be the single most important development in research communication of the past fifty years; it is hard to identify anything comparable. Its creator, Paul Ginsparg, received a McArthur Foundation Award in recognition of his achievement (*Science*, 298, 4 October 2002, p. 49).

The arXiv is an automated repository and distribution system for scholarly communication in physics, mathematics, nonlinear sciences, computer science, and quantitative biology, developed at the Los Alamos National Laboratory. The arXiv provides nearly comprehensive coverage of large areas of physics, and serves as an on-line seminar system for scholars in those areas. It contains more than a quarter million documents and boasts a user community of over 40,000 researchers. New submissions are received at a rate of more than 175 per weekday from scientists all over the world, 2/3 from outside the U.S., and the submission rate is increasing at about 9% per year.

Since its launch in 1991, the purpose of the arXiv has been to provide instant, equal, and uniform access to research materials on a global scale. ArXiv supports comprehensive aggregation, searching, and comparison in an open-access environment that is not possible in the traditional publisher-based environment.

Despite its powerful impact on scholarly communication in the sciences, arXiv complements, rather than replaces, formal, peer-reviewed publishing. The majority of submissions to arXiv are also submitted to conventional journals, where in areas like high-energy physics 70% plus are later published. A combination of heuristic screening mechanisms, scientists acting as moderators for various fields, and endorsement procedures contribute to ensuring, insofar as possible, that submissions to the arXiv are of refereeable quality. That is, they must satisfy minimum criteria to a degree that they would not be peremptorily rejected by a scholarly journal editor as manifestly inappropriate for publication. These mechanisms are an important component of why readers find the site so useful. Because of the quality of the content, active scholars in the represented fields are willing and eager to navigate the raw deposited material.

As a pure dissemination system, without the editorial functions associated with peer review, the arXiv operates at a fraction of the cost of a traditional scientific publisher. Many of arXiv's innovations going back to the early 1990s have been emulated by other online literature and database systems, both academic and commercial, and it is the forerunner of present open access experiments.

Researchers routinely testify to the critical importance of the arXiv to its user community, and statistics substantiate this. There were more than 20 million full-text downloads during the 2002 calendar year and an average of more than 300 full-text downloads of each submission in the seven years between 1996 and 2002. The usage is significantly higher than comparable online journals in the fields covered, and, most importantly, the access numbers have accelerated upward as additional conventional journals have come online over the past seven years. Usage per user has also increased over this period, signaling a measurable change in user behavior. The attraction of the arXiv continues to be its ability to provide "instant" dissemination of research, along with comprehensive aggregation of research across several related fields, extending back more than a decade.

When in 2001, Paul Ginsparg returned to Cornell, where he had done his graduate work twenty years earlier, he took-up a joint appointment in physics and in information science, and he brought the arXiv with him. Responsibility for daily administration of the arXiv came to the Library. Library staff are also engaged in developing new interfaces for both users and operators and further automating time-consuming processes. We are also documenting system operation and user support and articulating administrative policies that had remained relatively informal before. The transition has not been particularly easy. We are only now establishing the kind of stable operational environment with the kind of user service we view as essential, and administration still requires some active involvement by Ginsparg and a colleague who also transferred to Cornell from Los Alamos, Simeon Warner. I hope that we will have completed all elements of the transition by early-2005.

Presently, daily operation and continuing system development costs about $175,000 annually, not counting either Paul Ginsparg's nor Simeon Warner's time. During the first three years of the transition, the University Provost supported nearly ¾ of that cost, with the Library covering the remainder. However, the Provost's commitment will expire this fall, and full operational costs will be added to the Library's existing budget. In combination with system maintenance and upgrades and managerial and administrative support, costs of arXiv operation should average about $200,000 annually. Improved efficiencies will reduce the costs of daily administration, but extended scholarly moderation for the various fields and costs necessary to insure long-term preservation of content will contribute to increased costs over time.

If one thinks, however, of the scope of the benefits provided worldwide by the arXiv, these expenses are minimal, and Cornell is committed to sustaining its operation. So, why is arXiv indefinitely sustainable? Why did the National Science Foundation devote continuing support to its growth and development over a dozen years, and why will Cornell continue to insure its viability? First, its development is firmly based in the practices of the principal disciplines it serves.

The regular sharing of research in physics was already well underway via a range of different means when Paul Ginsparg envisioned a way to dramatically improve and expand those practices. Thus, it was highly compatible with the culture and needs of its clientele. At the same time, it did not necessitate change from other existing publications practices of its participants.

Additionally, its basic design offered economies of scale, allowing it to grow rapidly without significant increases in operational costs, resulting in a steadily improving cost/benefit ratio. Most importantly, however, as it gained mass, it became essential to its user community, embedded in basic practice. In combination with its operational efficiencies, it is the critical importance of the arXiv to its users and its compatibility with existing culture and practice that insures its sustainability. In his article, Don Waters describes the importance of this type of relationship with a user community as one in which digital resources have "such an impact on scholarship that their disappearance is not an option."

In spite of their very different operational models, both Euclid and the arXiv are explicitly designed in address the needs of particular user communities and to be compatible with their disciplinary cultures. ArXiv has attained necessary economies of scale, and it is clear that Euclid must achieve improved economies to attain financial stability and to achieve its principal goals.

Organizational Setting

In conclusion, I will address the question of appropriate setting for such endeavors. I will focus on the potential of research universities in fulfilling this role, but I recognize that there are other options, and Don Watters addresses some of the advantages of independent non-profit entities like JSTOR, ARTstor, and similar ventures being encouraged by Ithaka Harbors, Inc. I suspect that in spite of their long-term institutional stability and their close relationship with both the creators and users of scholarly information, Waters perhaps doubts that universities will maintain their focus on the actual business of publishing, or that they will make the organizational changes necessary to sustainable success. Nonetheless, the Mellon Foundation has encouraged innovative collaborations among university presses as a means of addressing the crises in monographic publishing and is now supporting publishing endeavors in several university libraries.

Although research libraries' historical role in acquiring and preserving the record of scholarly production is critical in addressing these issues, perhaps more important today is library operation of highly sophisticated production technologies. Libraries are also now well organized to ensure 24/7 system and user support. Yet, publishing is different. Libraries do not have a record of entrepreneurship, and issues of marketing, sales, cost accounting, and business planning are not traditional areas of strength. At Cornell, we have had to add staff with greater financial and publishing expertise. We have drawn on business

and marketing consultants when necessary. Most importantly, however, we have begun to think differently across a whole range of functions, enabling us to address new goals while drawing on a broad range of existing library strengths. At Cornell, our Digital Consulting and Production Services (DCAPS) integrates a range of services, including digitization, metadata services, copyright management, technology consulting, and electronic publishing, within a single service model. Though focused on providing an integrated solution for faculty, it is an enabling a suite of services supporting a broad range of activities on campus and beyond. This capacity to develop new synergies is essential in improving library operation, and it is providing us with the necessary managerial and financial management structure to address issues of sustainability in an effective manner.

Other critical elements are realistic planning and effective decision-making based on sound economics, the focus of today's workshop. In preparing for this workshop, I found an appropriate closing remark for today's presentation in the article that Zsuzsa Koltay and I published in 2001. It seemed obviously true then, at the end of the dot.com bust, and it certainly applies for us today as we seek to devise and implement new models for scholarly publishing, "Good ideas will not make up for bad economics." (Koltay and Hickerson, 2001, p. 94)