SUSTAINABILITY CASE STUDY: EXPLORING COMMUNITY-BASED BUSINESS MODELS FOR ARXIV

Oya Y. Rieger
Cornell University Library
Ithaca, NY
USA

Simeon Warner
Cornell University Library
Ithaca, NY
USA

ABSTRACT

arXiv.org is internationally acknowledged as a pioneering and successful open access digital archive for research articles. The case study discusses the efforts to establish a community-based sustainability strategy to ensure the longevity, effectiveness, and success of the service. It also describes the costs associated with running the repository that take into consideration both daily operational costs and efforts in improving its technical architecture and functionality.

1. INTRODUCTION

Started in August 1991, arXiv.org emerged as an exemplary disciplinary digital archive and open-access distribution service for research articles. The e-print repository has transformed the scholarly communication infrastructure of multiple fields of physics and plays an increasingly prominent role in a unified set of global resources for physics, mathematics, computer science, and related disciplines. It is firmly embedded in the research workflows of these subject domains and has changed the way in which scholarly articles are shared, making science more democratic and allowing for the rapid dissemination of scientific findings.

arXiv moved to Cornell in 2001, and is operated and maintained by Cornell University Library. The Library is committed to maintaining arXiv as an open access service, free to submitters and users alike. However, we believe that as a public good, arXiv should be supported by those institutions that use it the most. In an effort to address the long-term sustainability of this critical open access repository, the Library has developed a collaborative business model based on income generated by contributions from the institutions that are the heaviest users of arXiv [1].

In the first part of this case study we will describe the business planning process to provide a case study of opportunities and impediments in developing alternative business models. Financial stability alone is not sufficient to sustain a service such as arXiv, it must also be developed and improved to meet the needs and expectations of those who use it. Also critical is understanding the long-term preservation challenges associated with running a repository. In the second part of the case study we will outline our evolving technology plans including the underlying platform, preservation, services and interoperability.

2. SUSTAINABILITY: PROVIDING ENDURING ACCESS

Ithaka's 2008 report on sustainability provides a comprehensive review of a variety of business models for supporting online academic resources [2]. The report defines sustainability as "the ability to generate or gain access to the resources financial or otherwise needed to protect and increase the value of the content or service for those who use it." Therefore, keeping open access academic resources such as arXiv sustainable involves not only covering the operational costs but also continuing to enhance their value based on the needs of the user community. Furthermore, sustainability involves running a robust technical operation and addressing the digital preservation requirements of a system to ensure its long-term longevity and usability. Such a financial commitment is likely to be beyond a single institution's resources.

Scholars worldwide depend upon the stable operation and continued development of arXiv. Sustainability is best assured by aligning revenue sources with the constituents that realize value from arXiv, and by reducing dependence upon on Cornell University Library's budget. Our collaborative business model aims to engage the institutions that benefit from arXiv in defining the future of the service.

3. BUSINESS MODEL

3.1. Business Planning Process and Motivating Factors

arXiv moved to Cornell in the summer of 2001, and the Cornell University Library currently provides the bulk of arXiv's operating costs. Currently at $400,000 per year, the cost of operating arXiv is comparable to the entire physics and astronomy collections budget of the Cornell University Library. The business planning process began in June 2009 necessitated by the significant budget cuts that required the Library to review its programs and funding sources. Although the formal sustainability planning process was triggered by financial pressures, the Library was already engaged in exploring funding sources to support and further develop the archive, such as seeking an endowment and considering grant funding opportunities.

The first phase of the sustainability planning process involved a landscape analysis and a survey of arXiv stakeholders' positions and opinions on the e-print repository's future. Also critical during the assessment phase was expanding our understanding of the income models for open access and pros and cons of emerging practices. As we surveyed the positions of administrators, managers, and scholars from libraries and research centers, we often received the following concerns and questions:

- When is arXiv going to replace the formal journals?
- How will you address the free rider problem?
- Why not charge scholars per submission?
- What are the benefits for my institution?
- How will you structure a governance model?
- Are you opening a floodgate? Will other open access initiatives also start requiring contributions from the user community?
- What are the other potential sources of revenue?
- What is your long-term plan?

Based on a thorough review of available funding models [2, 3] and an extensive survey of arXiv stakeholders, we have considered many possible support options that are compatible with the Cornell University Library's mission. These include: sponsorship and advertising; donations; endowment; creation of "freemium" services; and support from funding bodies, scholarly and professional societies, and publishers. We consider the current plan as short-term and over the next three years will work with our advisors and supporting institutions to develop a long-term plan. The arXiv white paper further describes our planning process as well as addressing the questions raised by stakeholders during the input gathering process [1].

We also have considered the role of the Sponsoring Consortium for Open Access Publishing in Particle Physics (SCOAP3) initiative for our financial planning.[1] arXiv would potentially be a beneficiary of redirected funding administered by the SCOAP3 consortium. It is not clear, however, when this initiative will meet its funding goal. It should also be noted that SCOAP3 is restricted to High Energy Physics (HEP) and particle physics content only, which represents between 18% and 40% of submissions to arXiv (depending on how broadly the subject area is construed). Thus SCOAP3 could potentially subvent a similar fraction of arXiv's operating costs. It would be unreasonable to expect SCOAP3 or HEP labs to cover the entire cost of arXiv.

Another question we often received during our initial assessment process was the relationship between arXiv and other related databases such as Inspire and the Stanford Physics Information Retrieval System (SPIRES).[2] arXiv has long enjoyed close ties with the SPIRES and Inspire initiatives and project partners at the European Organization for Nuclear Research (CERN), SLAC, Deutsches Elektronen-Synchrotron (DESY), and Fermilab. We expect this collaboration to result in improved services for the HEP community, and we continue to investigate ways to share tools and software. However, the scope of INSPIRE is narrower than arXiv and we do not expect to find significant savings in operating costs. We continue to investigate additional partnerships that will enable us to improve the discovery and interoperability features of arXiv. We also see perceive such partnerships essential for bringing cost efficiencies.

3.2. Short-Term Business Plan

Currently, we are implementing an interim business arrangement (2010-2012) that aims to generate funds through recurring subsidies from the libraries at academic institutions, research centers, government laboratories, and other organizations that are the heaviest users of arXiv. The model entails a tiered structure of annual support requests similar to many other open-access funding models. The tier-based support structure is based on the previous calendar year's download activity and is applied equally to academic institutions, research centers, government labs, and other organizations. The 3-tiered institutional support model suggests institutional contributions within the range of $4,000 and $2,300 per year. We seek support from institutions representing the most active users of arXiv, in both the United States and other countries. Cornell University Library will continue to provide 15% of arXiv's operating budget, an amount many times higher than the support we will request from other heavy user institutions.

The calendar year 2010 budget for arXiv is $400,000, which includes costs for personnel and operating expenses [4]. The operational expenses include server ware, backups, storage, and preservation services. Staff salaries account for nearly 80% of total annual expenses. Running the repository involves 4.66 FTE staff, including user support, programming, system administration, and management. With over 60,000 new submissions per year one may think of this as an effective cost of approximately $7 per submission. Alternatively, with over 30,000,000 full-text downloads per year this is an effective cost of approximately1.4 cents per download.

We have no plans to impose article processing charges or submission fees. Barrier-free submission and use is

[1] http://scoap3.org/

[2] Inspire is a HEP information system that aims to integrate existing databases and repositories to host the entire corpus of the HEP literature worldwide. Run by the Stanford Linear Accelerator Center (SLAC), the SPIRES is a database of particle physics literature.

one of the founding principles of arXiv. We have considered requesting donations at time of submission but have concluded that such fundraising would incur greater overhead than the institutional support model, and would not engage our peer institutions. We also want to ensure broad international contributions to the repository without financial expectations from the authors. We are committed to maintaining arXiv as an open-access resource that anyone may use to download and read articles as well as allowing submissions free so that all appropriate articles can be accepted.

4. TECHNICAL PLANS FOR ENDURING ACCESS AND ADVANCING ARXIV

4.1. Scalable and Expandable Architecture for Sustainability

The arXiv software has been developed in-house over many years and this has both benefits and burdens associated with it. To keep arXiv sustainable, it is important to re-engineer the software to layer arXiv-specific functionality over generic repository software. Creating a generalized architecture will facilitate efficient technology management processes and allow the implementation of digital preservation procedures and policies.

The arXiv software was developed in-house at the Los Alamos National Laboratory and Cornell over the past eighteen years. The software has evolved and predated most other repository systems. It is predominantly written in Perl with components that use Java, PHP and Python. Metadata and user information is stored in a MySQL database and Lucene is used to provide the search service. The three server machines that provide the main arXiv.org site are supported by Cornell's central IT organization with 24x7 support. Mirror sites are locally supported and receive updates daily.

While the underlying technology has been periodically updated, the system requires significant internal re-engineering to support an evolving technological landscape, increased growth and use, and to ensure the sustainability of the service. The arXiv repository architecture includes some elements that are specialized to the user community, for instance the TeX processing system and the optimized administrative workflow in support of submissions and ingest. Other repository features are more generic and are good candidates for replacement with standard components in order to reduce costs and free developers for the development of new features and services.

We are in the process of surveying and assessing repository technologies. For example, one of the options is adopting a very standard platform such as Fedora for underlying repository functionality. Another strategy is implementing a community-based archival solution such as the Invenio system,[1] which is common within the physics community (developed at CERN). Such a system will lend itself for building features that target e-print archives and also will support the development of shared tools and web services that factor in disciplinary scholarly communication patterns and values.

The second element in our future technical plan is digital preservation of arXiv content in order to support the long-term maintenance of bitstreams and ensure that digital objects are usable (intact and readable), retaining all quantities of authenticity, accuracy, and functionality deemed to be essential when articles (and other associated materials) were ingested. Formats accepted by arXiv have been selected based on their archival value (TeX/LaTeX, PDF, HTML, OOXML) and the ability to process all source files is actively monitored. The underlying bits are protected by standard backup procedures at the Cornell campus and off-site backup facilities in New York City provide geographic redundancy. The complete content is replicated at our mirror sites around the world and additional managed tape backups are taken at Los Alamos National Laboratory.

The Cornell University Library is developing an archival repository (to be operational in May 2011) that will support preservation of critical content from institutional resources including arXiv. All arXiv documents, both in source and processed form, will be stored in this repository and there will be ongoing incremental ingest of new material. We expect that the preservation costs for arXiv will be borne by the Cornell University Library leveraging the archival infrastructure developed for the library system. As an interim solution and also a secondary archival strategy, we are also assessing community governed archives such as CLOCKSS and TRAC-certified services such as PORTICO.

4.2. Sustainability Through Innovation

Keeping open access academic resources such as arXiv sustainable involves not only covering the operational costs but also continuing to enhance their value based on the needs of the user community. arXiv's success has relied upon a highly efficient use of both author and administrative effort, and has served its large and ever-growing user base with only a fixed-size skeletal staff. In this respect, it long anticipated many of the current "Web 2.0" and social networking trends, providing a framework in which a community of users can deposit, share and annotate content. It also helped initiate an open access to scholarly literature movement and continues to play a leading role in such endeavors.

[1] CDS Invenio digital library system is a suite of applications that provides the framework and tools for building and managing e-print servers. The software is free and can be licensed under the GNU General Public Licence (GPL).

A critical aspect of the arXiv sustainability plan is enabling interoperability and creating efficiencies among repositories with related and complementary content to reduce duplicate efforts. Organizations with institutional repositories are usually keen to have them used, and would like to avoid the need to for authors to make multiple deposits. SWORD (Simple Web-service Offering Repository Deposit) aims to lower the barriers in contributing content in multiple repositories [5]. arXiv implemented the SWORD protocol for automated deposit over a year ago. This protocol enables both multiple deposits from a single tool and deposit from another repository. However, it has yet been used to address the 'multiple deposit problem' while it has been successfully used by journals and conference systems depositing in arXiv.

Digital data and associated multimedia information such as images and audio/video are becoming an integral part of scientific publications. To maintain its innovation role in scholarly communication, it is essential for arXiv to develop features in support of the deposit and archiving of supplementary information objects that are associated with a given paper. Also critical will be to factor in such multimedia content in the development of our preservation plans.

The Cornell University Library frequently receives requests to extend arXiv to include other subject areas. Due to limited resources, we have adopted a measured approach to expansion because there is significant organizational and administrative effort required both to create and to maintain new subject areas. Adding a new subject area involves exploring the user-base and use characteristics pertaining to the subject area, establishing the necessary advisory committees, and recruiting moderators. Although arXiv.org is the central portal for scientific communication in some disciplines, it is neither feasible nor necessarily desirable to play that role in all disciplines. However, arXiv can provide a model for other communities through improved service to its existing dedicated user communities, and act as an essential component of a global networked scholarly communication system. We anticipate that system will become increasingly broad in its subject area coverage, and increasingly diverse in its component databases, repositories, and other online tools and services.

5. PLANS FOR DEVELOPING A LONG-TERM SUSTAINABILITY STRATEGY

We realize that our business model needs to be responsive to the shifting ecology of scholarly publishing. Our current sustainability model represents a short-term strategy for the next three years. We are collaborating with the heaviest user institutions in the US and abroad in our effort to reposition arXiv as a vested online scholarly resource, an asset with shared benefits and accountability. We are very pleased that so many institutions have already stepped forward to share the cost of arXiv. As of June 2010, 79 institutions have pledged their support totaling to $283,000 in contributions. Complementing this short-term business planning efforts, we formed an international advisory group, which will provide an essential consultative role in developing diverse sustainability strategies for this critical international resource.

Over the next few years we will develop a long-term business plan that provides a strategic framework to protect and increase the value of arXiv for those who use it. Ideally this will comprise a blend of ongoing underwriting from Cornell University Library and support from the academic library community and research centers. It might also include support from scholarly societies, an endowment, or funding agencies such as the NSF. We will strengthen existing collaborations (e.g. with the INSPIRE project of CERN, SLAC, DESY and Fermilab) and develop additional partnerships that allow arXiv to provide better services or to share the support burden. Advice from the sustainability advisory group and other supporting institutions will be used in developing this long-term business plan. Also critical for our long-term strategy is developing an approach for reliable and committed stewardship in order to sustain the technical operation and innovative track record of this highly valued repository.

6. CONCLUDING REMARKS

The arXiv case study presented in this paper illustrates the need to approach digital preservation of repositories holistically by taking into consideration a range of lifecycle and usability issues as well as factoring in the changing patterns and modes of scholarly communication. As we collectively address the creation and management of community-based infrastructures, we need to factor in financial needs, usability factors, innovation in discovery and access, and enduring access. arXiv complements, rather than competes with, the commercial and scholarly society journal publishing market. A critical question for the repository and preservation community to address is the versioning of scholarly articles, from initial submission to pre-print archives to their final publishing in formal scholarly journals.

One of the goals of our business planning initiative is to provide a case study that can be used by other institutions with similar repository responsibilities. As support for open access publishing increases and the reliance of users on free resources grows, it is inevitable that educational and cultural institutions will need to collaborate in experimenting with different funding strategies. What is essential is for organizations with such undertakings to share their experiences and lessons learnt with the broader community in order to collectively enhance our understanding of issues and pros and cons of potential strategies. To this end, Cornell

University Library is committed to continue discussing the sustainability planning process and outcomes with our colleague through different forums and channels.

7. REFERENCES

[1] arXiv Business Model White Paper, January 2010. http://arxiv.org/help/support/whitepaper.

[2] Guthrie, K., Griffiths, R., Maron, N. Sustainability and Revenue Models for Online Academic Resources. An Ithaka Report. 2008. http://www.ithaka.org/ithaka-s-r/strategy/sustainability-and-revenue-models-for-online-academic-resources

[3] Raym Crow. Income Models for Open Access: An Overview of Current Practice, 2009. http://www.arl.org/sparc/publisher/incomemodels/

[4] aXiv 2010 Budget, June 2010, http://arxiv.org/help/support/2010_budget

[5] SWORD (Simple Web-service Offering Repository Deposit) Deposit Lifecycle White Paper. http://www.swordapp.org/.