Key Issue
Subject and institutional archives: comparing the examples of arXiv and Cornell's institutional repository

OYA Y RIEGER
Associate University Librarian
Digital Scholarship Services
Cornell University Library
USA

A few months ago, while I was participating in a conference about open access infrastructures, a delegate from a governmental agency asked, "Why does each library need to maintain a repository for their own scientists?" He was rightfully wondering if a broad collaboration in building a network of archives will provide a durable and extensible technology and service framework for ever-increasing digital scholarly content.[1] The ensuing discussion did not offer a plausible response but accentuated that we do not have in place a plan for building an expandable infrastructure to facilitate communication and exchange of information among rapidly proliferating distinct instances of institutional and subject repositories.

Over the past two decades, open access digital repositories have become an increasingly vital component of the scholarly communication infrastructure. Such repositories are expected to facilitate broad and unrestricted discovery of information and ensure enduring access to knowledge through preservation. The nascent repository landscape is heterogeneous, featuring different repository technologies, content types, functionalities and user communities. Although sharing and interoperability have been core library values, the vision of creating a shared repository infrastructure continues to be an elusive one. As demonstrated by the 2008 Subject and Institutional Repositories Interaction Study, while repository managers express great interest in interacting with other archives, in practice repository visions are often driven by institutional priorities and other local factors.

"Although sharing and interoperability have been core library values, the vision of creating a shared repository infrastructure continues to be an elusive one."

Similar to its peers, Cornell University Library (CUL) in the United States of America has had an active repository agenda for open access content and maintains several such systems. The most eminent one is arXiv, a subject repository. The Library also operates an institutional repository called eCommons. I will characterize and contrast these two archives from the perspectives of origin, content, user community and sustainability in order to illustrate the impediments and virtues associated with subject and institutional repositories (IRs).

eCommons

Origin

eCommons is Cornell's institutional repository with the mission to 'provide long-term access to a broad range of Cornell-related digital content of enduring value'. It was launched in 2002 by the Cornell University Library. It was motivated by the vision of the Dean of the Faculty to create 'an economical vehicle for openly-shared access to formerly inaccessible but intellectually-rich digital resources'. This mandate dovetailed with CUL's interest in building a repository reflecting the academic priorities and significant research areas of the University.

Content

The institutional repository has a set of policies to define its collection, deposit, access, withdrawal, alternation, privacy and preservation.[2] However, it is not a curated collection and does not have any quality control mechanisms for deposited content. It is composed of a range of documents including articles, books, reports, slides, theses, dissertations, preprints, visual images, data sets, course materials and AV resources – whatever the Cornell community members find appropriate for depositing. A fairly broad range of Cornell academic areas are represented in the repository but without a significant collection, except for the electronic theses and dissertations deposited by the Graduate School.

User community

Similar to many other IRs, eCommons garners a modest use and, since its inception in 2002, has accumulated approximately 20,000 items. Annual downloads are in the vicinity of one million hits. Both items added and downloads have seen very small annual increases. The Library mediates most submissions to eCommons. For instance, in 2009, over 86% of all submissions to eCommons were carried out by Library staff. Content in eCommons is certainly visible to a worldwide audience. Roughly 46% of all eCommons sessions in 2009 were initiated by users outside the US. Overall, eCommons has only partially fulfilled its original intent. We had a vision of providing a way to manage and preserve Cornell's digital academic assets to enable greater visibility and accessibility over time. However, the idea of an institutional repository does not have a strong traction with faculty, who often feel closer to their own disciplinary networks.[3]

" … the idea of an institutional repository does not have a strong traction with faculty, who often feel closer to their own disciplinary networks."

Sustainability

eCommons is based on DSpace and is maintained by the Library's IT department. Initial funding was provided by the Atlantic Foundation and operational responsibility transferred to CUL in 2008. It is well integrated into the Library's digital infrastructure and therefore is seen as a core library service without any concerns about its sustainability. In 2009, the operational budget was $129,000, including staff and server support: in other words, $56.82 per submission and 16.5¢ per download.

arXiv

Origin

Founded by Paul Ginsparg in August 1991, arXiv was originally developed to supersede an international e-mail distribution list for physics preprints that was manually operated. It was originally hosted at the Los Alamos National Laboratory and called the LANL preprint archive.

Content

arXiv is the primary daily information source for hundreds of thousands of researchers in physics, and plays an increasingly prominent role in mathematics, computer science and other related fields. It provides an instant communication mechanism for scientists and complements the formal publishing process, which may take several months. Unlike eCommons, it is a moderated repository. Submissions are reviewed by expert moderators to verify that they follow accepted standards of scholarly communication. Additionally, an endorsement system uses community feedback to pre-screen new submitters. Enabling interoperability and creating efficiencies among repositories with related and complementary content has been a key priority for the arXiv team. For instance, SWORD protocol enables both multiple deposits from a single tool and deposits from another repository[4]. Although it has not fully solved the 'multiple deposit problem,' it has been successfully used by journals and conference systems depositing in arXiv.

User community

Through Paul Ginsparg's leadership, the service has been informed by the disciplinary cultures represented in the digital repository. The submissions are screened by volunteer subject-specific moderators to ensure content is relevant to current research in the specified disciplines. arXiv has facilities to harvest and display references and links to formally published versions of articles based on the deposited e-prints, thus providing an overt link to peer review. arXiv currently includes over 700,000 e-prints and is visited by 400,000 distinct users per week.

"arXiv currently includes over 700,000 e-prints and is visited by 400,000 distinct users per week."

Sustainability

Since 2001, the service has been operated by Cornell University Library. In January 2010, Cornell has established a voluntary institutional contribution model and invited pledges from the top 200 libraries and research laboratories accounting for more than 75 percent of annual institutional downloads[5]. This was based on the fact that only 0.5-0.7 percent of use is from the Cornell community while the Library was covering the entire costs. The community-funding model entails a tiered structure of annual support requests ($4,000 to $2,300 per year). Based on a budget of $330,000 and 40 million paper downloads for 2010, each e-print costs merely 0.08 cents per download and the cost per submission is $4.70.

Conclusion

The existing repository ecology has complex architectures and features that are optimized to fulfill the specific needs of institutional, subject or archival repositories. The landscape is becoming even more heterogeneous with the addition of scientific social networking sites that profile local scholarly activities and open data initiatives that focus on data curation models. As we plan the future of repositories, especially how they communicate with each other, we need to factor in the following aspects:

- interoperability arrangements that link a given repository to related systems, services and communities
- versioning of scholarly articles, tracking them from initial submission to preprint archive to final publication in a formal scholarly journal ensuring the authority and integrity of e-prints and distinguishing between succeeding versions, such as a pre-print article and its published version in a scholarly journal
- features that support supplementary information objects such as underlying data, auxiliary multimedia content and research methodologies
- functionality and arrangements that lower barriers to contributing content to multiple complementary repositories.

In quest of a seamless discovery environment, we need to link the burgeoning corpus of institutional repositories with related subject systems in order to support version control as well as create a critical mass of related materials on particular topics. There is a great potential for subject and institutional repositories to function in a complementary fashion by leveraging their particular strengths.[6] There are standards, technologies, practices and policies (either ready or in development) that would allow such synergies; however, there is need for a broad architectural map to conceptualize such an information environment. Achieving an integrated vision is becoming even more urgent with the increasing open access polices for publicly-funded research and institutional open access mandates.

"There is a great potential for subject and institutional repositories to function in a complementary fashion by leveraging their particular strengths."

References and notes

1. Terms such as archives and repositories continue to be used interchangeably, sometimes depending on the preferences of specific communities: http://ecommons.library.cornell.edu/policy.html#content (accessed 25 January 2012).
2. Rieger, O Y, Opening Up Institutional Repositories: Social Construction of Innovation in Scholarly Communication, *Journal of Electronic Publishing*, 2008, 11(3). DOI: http://dx.doi.org/10.3998/3336451.0011.301 (accessed 25 January 2012).
3. Warner, S, SWORD V1 Case Study – arXiv, 2008: http://swordapp.org/sword-v1/sword-v1-case-studies/sword-v1-case-study-arxiv/ (accessed 25 January 2012).
4. Rieger, O Y, Assessing the Value of Open Access Information Systems: Making a Case for Community-Based Sustainability Models, *Journal of Library Administration*, 2011, 51(5-6), 485–506.
5. Darby, R M, Jones, C M, Gilbert, L D and Lambert, S C, Increasing the productivity of interactions between subject and institutional repositories, *New Review of Information Networking*, 2008, 14(2), 117–135: DOI: http://dx.doi.org/10.1080/13614570903359381 (accessed 25 January 2012).

E-mail: oyr1@cornell.edu

To cite this Key Issue:

Rieger, O Y, Subject and institutional archives: comparing the examples of arXiv and Cornell's institutional repository, *Insights*, 2012, 25(1), 103–106, doi: 10.1629/2048-7754.25.1.103

To link to this Key Issue:

http://dx.doi.org/10.1629/2048-7754.25.1.103