

INSTITUTIONAL REPOSITORIES

A White Paper for the UNC-Chapel Hill Scholarly Communications Convocation January 2005

by
Wallace McLendon, Deputy Director, Health Sciences Library

Introduction

The institutional repository (IR) concept has gained momentum as universities begin to question the logic of buying back its research, as libraries drop journal subscriptions due to publisher fees outstripping resources, and as taxpayers question paying for research twice by funding the research itself followed by purchasing journal subscriptions to discover the research findings. IRs can preserve and provide access to a university's unpublished material, establish alternatives to the high costs of traditional publications, and contribute to a university's prestige. As information and knowledge resources are increasingly digitized and distributed by local and global networks, those facing the above issues are exploring alternatives to the preservation and distribution of information.⁽¹⁾

Definitions

IRs manage and create supporting services to store, preserve, and disseminate an organization's digital information or knowledge assets created by faculty, research staff, and students with few if any submission or access barriers.⁽¹⁻³⁾ While university archives manage administrative records to satisfy mandates and preserve materials pertaining to the institution's history, archivists outside of legal mandates exercise broad discretion in determining which papers and other digital objects to collect and store. IRs tend to contain any product generated by the institution's students, faculty, non-faculty researchers, and staff and includes student electronic portfolios, classroom teaching materials, the institution's annual reports, video recordings, computer programs, data sets, photographs, and art works—virtually any digital material that the institution wishes to preserve.⁽³⁾ Some initial repositories are and will be discipline-based and populated by faculty from multiple universities (e.g. <http://arxiv.org/>).⁽⁴⁾ Disciplinary repositories are usually preprints or e-prints of scholarly articles and technical reports, while an institutional IR is eclectic with assets drawn from the institution's diverse teaching and research output.⁽¹⁾

History

Repositories began with man's first storing and protection of artifacts and information succeeded by the formalization of those efforts through libraries and museums. In 1988, Peter Drucker's 1988 *Harvard Business Review* article "The Coming of the New Organization" declared that an organization's knowledge was its most important asset and to manage that asset well was to ensure the organization's success.⁽⁵⁾ Thus began the knowledge management movement of the 1990's that reached beyond book and article

“containers” and placed value on all knowledge explicit and tacit, in datasets and graphics, in e-mails and sketches. By 2000, it was becoming easier for individuals and groups to create and disseminate content using desktop tools and networking which challenged universities to coordinate, share, and preserve its digital assets.⁽¹⁾ In 2002, two seminal events occurred when the Massachusetts Institute of Technology (MIT) collaborated with Hewlett-Packard Corporation to launch an open-source institutional repository entitled DSpace and the Scholarly Publishing and Academic Resources Coalition (SPARC) published, “The Case for IRs: A SPARC Position Paper.”^(6,7) From DSpace emerged a new strategy for universities to capture their creativity and research as well as pose an alternative to the high-costs of scholarly communication. In 2003, with funding from The Andrew W. Mellon Foundation and other sources, MIT's DSpace was replicated and the software released under an open source arrangement, greatly lowering cost and expediting development. While the MIT software is not the only option available (e.g., University of Southampton in the U.K. <[http:// www.eprints.org/](http://www.eprints.org/)>), it has become the most general-purpose.⁽²⁾ The synergy of the Internet, the decrease in online storage costs, and the development of standards set the stage for Institutional Repository experimentation and eventual implementation.

Current Projects

An up-to-date inventory of university IRs can be found at <http://archives.eprints.org/>. As cited earlier, MIT has provided leadership in the implementation of DSpace. Numerous universities including Cornell, Cranfield, Drexel, Erasmus, Ghent UGent, Groningen, Hong Kong, Indiana, Kansas, Maryland, Minho, the Netherlands, North Carolina at Chapel Hill SILS, Ohio, Oregon, Parma, Purdue, Toronto, Roskilde, Vrije, and Washington School of Medicine have adopted the DSpace institutional repository platform.

GNU EPrints Archive Software (<http://software.eprints.org/>), an open source software suite, is used by the Universities of Australia, Bath, Beijing, Bologna, Caltech, Cape Town, Catholique de Louvain, Central European, Connecticut State, Dublin City, Duke Law Faculty, Durham, Edinburgh, Firenze, Goizueta, Gotenborg, Hokkaido, Indiana Biology Department, Iowa State, Johns Hopkins Latin American Scholars, Ljubljana, Ludwig-Maximilians-Universitat, Lund, Melbourne, Minho, Monash, Montreal, Munchen, National of Ireland, Oxford, Paris X Nanterre, Quebec a Montreal, Queensland, St. Andrews, Sao Paulo Institute Tilburg, Simon Fraser, SMU Central, Studi di Padova, Stuttgart, Tasmania, Toronto Scarborough Campus, Trento, Vaxjo, Virginia Tech, Uppsala, Warwick, Windsor, York and several institutes of technology including the Curtin, New Jersey, Paris, Swedish, and Indian.

Multi-institutional projects include the ARNO project (Academic Research in the Netherlands Online), the DARE project (Digital Academic Repositories) a collective initiative by the Dutch universities, the eScholarship initiative of the California Digital Library to serve University of California units, the FAIR project funded by the UK Joint Information Systems Committee involving over 50 UK universities, and the Knowledge Bank at Ohio State University emerging out of the University's Distance Learning &

Continuing Education Committee to include all digital assets and information services available to the OSU community.

IRs and Open Access

“Open access” and IRs are symbiotic and the distinction between the two is frequently blurred. An institutional repository can be thought of as a container and open access as policies and standards (Open Archive Initiative and Open Archives Metadata Harvesting Protocol) that govern how IRs function. Both the open access and institutional repository movement grew out of the scientific and library communities' experience with disciplinary e-print archives or repositories.⁽⁸⁾ The genesis of the e-print repository is arXiv begun by Paul Ginsparg of the Los Alamos National Laboratory in 1991 and is now housed at Cornell. Containing 230,000 papers in the fields of physics, mathematics, nonlinear science, and computer science, the repository has always provided free access via the Internet. Such efforts have inspired the Budapest Open Access Initiative and the Public Library of Science.⁽¹⁾

Content

The content of an institutional repository is not a content free-for-all. IRs will be institutionally defined, scholarly, cumulative, perpetual, open, interoperable and may include pre-prints, peer-reviewed articles, monographs, enduring teaching materials, data sets, ancillary research material, conference papers, electronic theses and dissertations, electronic portfolios, the institution's annual reports, video recordings, computer programs, photographs, art works, etc.⁽³⁾ Perpetuity establishes that items submitted can not be withdrawn, format standards must be established, issues surrounding pre-publication materials deposited must be addressed, and participants defined.⁽³⁾

Whereas libraries practice selectivity based on lasting value and format, IRs will require expanded formats, additional resources, and greater inclusion. The scope of a repository's content will have to be defined and who defines that scope will need to be determined. While digital archives management (DAM) systems technology can handle a variety of formats, there are still limitations and a lack of standards that challenges metadata preparation and preservation. Technology capabilities and metadata standards, therefore, may influence content.⁽¹⁾ Copyright and licensing will dictate content as well. Authors publishing in traditional venues will work out agreements to house their publications within a university's repository.⁽⁷⁾

Intellectual Property

Intellectual property rights will impact repositories. Powerful capabilities to copy, reuse, repurpose and restrict content in repositories will increase ownership issues. Non-traditional formats such as e-prints, courseware, and a student's e-portfolio may have previously escaped university scrutiny and resulting policies. An educational component alerting all parties in the use and ownership of these new formats will be required.

Cornell, Brigham Young University, and MIT's Dspace program have model property and copyright polices.^(9,10,6)

Administration

Repositories can be owned by an individual, a group, an institution, a commercial organization, a consortium, or by government and defined along organizational, political, or constituency jurisdictional lines. The owner's major challenge is to balance ease of access and participation while ensuring mechanisms in place are consistent, scaleable, and robust. In planning IRs, the University as owner will need to acknowledge the strengths and limitations of existing units in the organization to manage this rich information environment. Branin suggests that the responsibility fall to a combination of entities including campus libraries and information technology.⁽¹⁾

Policies, processes, and technical infrastructure will be needed to manage submissions, versions, access, and system updates from departments, libraries, research centers, labs, institutes, and individuals. Some systems allow the embargo of submissions until they are reviewed based on institutional policies. The institutional repository manager will face a cumulative wave of submissions that will result in millions of digital objects and terabytes of storage. But universities do not have to face this daunting task alone. Existing consortia can pool resources and expertise resulting in economies of scale.⁽³⁾

The university will decide the degree of assistance – multimedia production, design, digitization, and metadata training and preparation – to provide authors. Self-archiving, templates, and automated services that delegate submission processes to the author may be emphasized. Indexing and metadata preparation will be at the heart of the repository's usability and interoperability. The university will also decide the level of support and access it will provide content consumers. As in all systems, managers will address a myriad of maintenance issues including backup systems, disaster preparedness, well timed migrations, etc.⁽¹⁾

Branin believes the ultimate challenge will be faculty and student participation. He does not see universities wielding the control corporate management has in requiring employees to deposit knowledge assets into a repository. He sites two methods that have been successful in early adopting universities: (1) conducting an inventory of current campus projects on campus, engaging participants concerning their experiences, and involving them in the planning of a campus institutional repository, and (2) identifying "communities of practice" as an organizational focus in implementing repositories.⁽¹⁾

Technology and Infrastructure

The Open Archival Information System (OAIS) Reference Model, Open Archives Metadata Harvesting Protocol (OAI-PMH), Metadata Encoding and Transmission Standard (METS), Shareable Courseware Object Reference Model (SCRORM), Publishing Requirements for Industry Standard Metadata (PRISM), Dspace, ePrints, FEDORA, bepress, Documentum, CONTENTdm, IBM's Content Management, and

Artesia's TEAMS offer standards and guidance on technical architectural issues.⁽¹⁾ There are commercially available digital asset management (DAM) systems in two categories: commercial turnkey systems and nonprofit open-source systems. Dspace and ePrints – nonprofit open-source systems – initial costs are low but require local development and support while commercial turnkey systems have higher start-up costs but promote lower long term expense. Dspace not only provides system software but shares carefully documented policies and procedures.⁽¹⁾ In addition to Dspace, there are two other free software packages developed for use with IRs: GNU EPrints from Southampton, UK, and CDSware from CERN, Switzerland. The Open Society Institute has published a free Guide to Institutional Repository Software.⁽⁷⁾ The university also may consider whether or not to implement searching and indexing functionality and just maintain and expose metadata, allowing other services to harvest content which lowers cost and support barriers for many institutions. This approach requires a file system to hold content and the ability to create and share metadata with external systems.⁽³⁾

Costs

Some early adopters report that the initial costs are modest in that most university research and content is already digital; however, resolving policy and standards requires intense investments in labor. Costs should recede as universities gain experience and establish routines.⁽⁷⁾ The establishment of IRs or a federation of them will be an expense in addition to fees paid to traditional publishing entities. Cost savings resulting from alternative content management and distribution may be realized years into the future. From this perspective, universities may need to clarify their motives for entering into such a venture. Universities may choose to be selective in their IR participation while outsourcing the more challenging content formats with companies like ProQuest or BioMed Central.⁽¹¹⁾ The creation of a five to seven year budget should be part of the institutional planning process as a means of measuring return on investment. To date, universities are finding that organizational costs involving policies, management and marketing outweigh the cost of the required technology.⁽³⁾

The California Institute of Technology has kept its cost low by integrating repository and library services while using free software such as ETD-db from Virginia Tech and Eprints from the University of Southampton.⁽¹²⁾ Dspace has cost Hewlett Packard and MIT approximately \$2 million with estimates that post development costs will require \$285,000 annually. A Dspace budget is available at its website. In 2003, Ohio State University invested \$265,000 to implement its Knowledge Bank.⁽¹⁾ Compared to total university budgets, MIT and OSU's repository program represents 1-2%.⁽¹⁾

Promise and potential

Crow and Branin describe the promise of IRs: efficient and cost-effective scholarly communication systems, good publicity for a university's research and teaching programs, effective mechanism for storing and accessing existing stored departmental and institutional documents, broader distribution of academic work via the internet, greater exposure of an author's work, less administrative burden on reporting

publications for research assessment and review exercises, central archives for an author's work, broader faculty exposure, ability to handle significant increases in the overall volume of research, alternative to escalating journal pricing, interoperability enabling interdisciplinary research, prestige shifted from publishers to universities, preserves the essential components (establishes intellectual priority, certifies quality of the research and/or validity of the claimed finding, makes the information accessible through dissemination, provides the ability for other researchers to become aware of new information, archives and preserves intellectual heritage) scholarly communication, improved discovery through simultaneous housing of content in multiple locations, places preservation of electronic content in the hands of professionally trained librarians rather than commercial entities that are being bought, sold and merged resulting in unpredictable content stewardship. ^(7,1) Moreover, universities are becoming aware that most of the direct labor and indirect cost are borne by the university. Publisher contributions have centered around distribution which has been answered by networks and the Internet. As components of traditional publishing are separated, registration, certification, and awareness functions can be handled by any organization with intellectual prestige, standing, and market position. ⁽¹⁾

Concerns

Lynch lists concerns about IRs including the entrenched investments by various parties involving the role prestigious journals play in the tenure process, large publishers have no incentive to change where they hold a monopoly, IRs may control where faculty or student works are deposited, rigid gate keeping like those in scholarly publications will discourage innovation, viewing IRs as the answer to existing scholarly publishing problems may create inappropriate policy constraints, IRs provide a much broader spectrum of new scholarly communication and to view them as just an alternative to traditional publishing sells the concept short, IRs should encourage participation through ease of use, IRs should not be viewed as a challenge or alternative to disciplinary repositories, IRs should be viewed as complementing existing venues of scholarly publication, universities should check the normal reflex to do what is popular and fashionable and avoid hasty Implementation, universities should realize they are taking risks and creating expectations, IRs represent an ongoing budgetary obligation, faculty trust and dependency go hand-in-hand, failure in the IR system can not be tolerated and would undermine broad social support for higher education, and universities will have to resolve how to handle the movement of faculty and their creative products from university to university. ⁽²⁾

Publisher concerns over IRs are exemplified in their recent response to a proposal before the US Senate that NIH funded research be deposited, after a six month embargo, in the PubMedCentral repository, thus providing free access. Representatives of the Professional Scholarly Publishing (PSP) and Association of American publishers have claimed that such a process would cause irrevocable disruption of scholarly discourse; cancellation of subscriptions; give publishers little choice but to enact author fees; undermine the economic foundation of established journals; establish a dangerous precedent by limiting an author's freedom to publish how, when, and where he or she

chooses; compromise the integrity of the scientific record; fail to acknowledge the value added by publishers; have an unknown impact on scientific and medical publications and professional societies that rely on subscription income for operations funding; and ignore publishers' investment of millions of dollars in digitizing their content.⁽¹³⁾

Future

Universities by definition create and store information in some form of repository already and may choose to broaden their capacities on individual campuses or through consortia. The economies of group participation will make IRs part of future consortia negotiations. Allowing researchers to search across IRs through the establishment of and adherence to standards is powerful enough to re-shape scholarly research if not societal norms. The field of IRs may be one mixed with public and private initiatives as non- and for-profit organizations refine their roles. Some outcomes are predictable while others are undoubtedly unforeseen. Regardless, scholarly communication will not look the same in a decade whether all universities choose to participate or not.⁽²⁾

Summary

University IRs will succeed if administrators, faculty and students share values surrounding knowledge asset management. The campus content creators must trust the university with their content and the university will need to win that trust through thorough planning, the establishment of realistic expectations, and continuous education.⁽¹⁾ In the near future, faculty and students will have choices to make in how they store and distribute their intellectual property and the university must position itself to be a viable choice. Whether or not IRs can change the monopolistic and pricing behavior of journal publishers is yet to be seen, but well thought out competing strategies and successful IR implementation will only enrich the ability of researchers to access a broader range of information.⁽⁷⁾ If Peter Drucker is correct in his assessment that an organization's value is based on its collective knowledge, the investment in institutional repositories should enhance the university's value to society.⁽³⁾

References

- (1) Brainin, Joseph. Institutional Repositories. Encyclopedia of Library and Information Science. Marcel Dekker, 2004, pp 1-11. <http://www.dekker.com/servlet/product/DOI/101081EELIS120020335> (accessed November 2004).
- (2) Lynch, Clifford A. Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL Bimonthly Report, 226, February 2003.
- (3) Johnson, Richard K. Institutional Repositories: Partnering with Faculty to Enhance Scholarly Communication. D-Lib Magazine, November 2002.
- (4) Report from the CIC Summit on Scholarly Communication: Access to Journal Literature, Executive Summary & Recommendations for Action. Chicago, Illinois. October 28/29, 2004.
- (5) Drucker, Peter. The Coming of the New Organization. Harvard Business Review

- on Knowledge Management. Harvard Business School Press, Cambridge, MA, 1998.
- (6) Dspace Federation <http://www.dspace.org/> (accessed November 2004).
 - (7) Crow, Raym, The Case for IRs: A SPARC Position Paper. The Scholarly Publishing & Academic Resources Coalition, 2002.
<http://www.arl.org/sparc/IR/ir.html> (accessed November 2004).
 - (8) Open Archives Initiative, <http://www.openarchives.org/> (accessed November 2004).
 - (9) Cornell University Copyright Policy, <http://www.research.cornell.edu/CRF/Policies/Copyright.html> (accessed November 2004).
 - (10) Brigham Young University Intellectual Property Policy, <http://ipsinfo.byu.edu/ippolicyt.htm> (accessed November 2004).
 - (11) Chillingworth, Mark. OA publisher launches new service that builds and maintains IRs. Information World Review, 15, Sep 2004.
<http://www.whatpc.co.uk/news/1158110> (accessed November 2004).
 - (12) Caltech Collection of Open Digital Archives (CODA), <http://library.caltech.edu/digital/default.htm> (accessed November 2004).
 - (13) Professional Scholarly Publishing (PSP) and Association of American publishers (AAP) <http://grants1.nih.gov/grants/guide/notice-files/NOT-OD-04-064.html> (accessed November 2004).