Introduction: Institutional Repositories: Current State and Future

Sarah L. Shreeves and Melissa H. Cragin

Abstract

Institutional repositories (IRs) currently exist in a rapidly shifting landscape without a clear consensus on their role in the academic environment. Low self-archiving rates have dampened hopes that IRs would have an impact on scholarly publishing models. Preservation programs, a stated goal of many IRs, are often not well established. In many cases, IRs are not part of a larger vision for services the library can provide to the institution, but are isolated projects without a strong base of support. Institutions are beginning to explore the role of IRs in the collection of materials like data sets. Given this environment, where will IRs be in the next five or ten years? This issue of *Library Trends* contains an impressive slate of articles from prominent practitioners and researchers in the field, who offer a range of perspectives on the current state of IRs in academic institutions and reflections on their future.

Introduction

This issue of *Library Trends* explores the current environment and possible future of institutional repositories. What is an institutional repository (IR)? An IR is a set of services and technologies that provide the means to collect, manage, provide access to, disseminate, and preserve digital materials produced at an institution. While most institutional repositories are based at colleges and universities, they also exist in governmental agencies, museums, corporations, and other organizations. Within colleges and universities, most IRs are managed by the library (Markey, Rieh, St. Jean, Kim, & Yakel, 2007). An institutional repository—as we refer to it here—is not only technological infrastructure in the form of software, servers, and storage, but the programs and resources that surround the

LIBRARY TRENDS, Vol. 57, No. 2, Fall 2008 ("Institutional Repositories: Current State and Future," edited by Sarah L. Shreeves and Melissa H. Cragin), pp. 89–97

technological infrastructure. As of October 2008, OpenDOAR[1] lists over 1,000 institutional repositories from around the world. These range in size from just a few items to close to 200,000 (DSpace@Cambridge [http://www.dspace.cam.ac.uk/]).

The goals for an institutional repository can vary widely. When IRs were first promoted in the early part of this decade, with the introduction of Eprints from the University of Southampton in 2000 and DSpace from the Massachusetts Institute of Technology and Hewlett Packard in 2002, the focus tended to be on providing open access to research and scholarship produced at an institution. With the focus on open access, many saw IRs as a way for libraries to push back against the serials crisis (Prosser, 2003). Others saw them as a means to promote the institution by showcasing its research and scholarship, or as a means to provide management and preservation of research and other material produced at an institution. The following quotes from three early papers on IRs or open access illustrate these varying motivations. In an influential early essay, Clifford Lynch (2003) describes institutional repositories as: "essentially an organizational commitment to the stewardship of . . . digital materials, including long-term preservation where appropriate, as well as organization and access or distribution." In 2004 Harnad et al. write: "But the self-archiving method with the greatest potential to provide OA [open access] is self-archiving in the author's own university's OAI [Open Archives Initiative]-compliant Eprint Archives . . ." (p. 312). In 2002, Raym Crow wrote in a SPARC position paper: "While institutional repositories centralize, preserve, and make accessible an institution's intellectual capital, at the same time they will form part of a global system of distributed, interoperable repositories that provides the foundation for a new disaggregated model of scholarly publishing" (p. 6). So are IRs meant to exert economic and social pressure on the current scholarly publishing model—and publishers—by supporting open access to published research? Are they primarily a vehicle for open access to research? Alternatively, are they primarily for the stewardship of material—such as grey literature, administrative records, and data sets[2]—that may be at risk within an organization? Or, should they support both purposes, as Crow (2002) suggests? Even now, eight or so years after institutional repositories were first introduced in the United Kingdom and United States, the motivation for IRs appears to be unclear. Abby Smith notes in her foreword to the Census of Institutional Repositories that there is still no clear consensus around what purpose IRs serve (Markey, Rieh, St. Jean, Kim, & Yakel, 2007, p. ix). Thomas and MacDonald (2007) also note the difficulty of categorizing repositories generally.

Just as there is a range of motivations driving the implementation of IRs, the type of content contained in repositories can also vary; this variation is often dependent, of course, upon the goal of the repository. Content can include: published articles, conference papers and posters,

book chapters, preprints, technical reports, working papers, presentations, data sets, websites, dissertations, theses, and other student work, digitized material from library holdings, administrative records, curricular materials, audio, video, and other materials. The range of content in IRs poses its own set of problems. McDowell (2007) estimates that only 13 percent of the content of institutional repositories are published, peer reviewed items. Nonpublished material, grey literature, and data sets may fit well into an institutional repository's collection policy, but inclusion of these types of material has sometimes brought IRs and repository managers under attack for diverting focus from the goals of the open access movement (Poynder 2006). The sheer range and inconsistency of collections can cause confusion for users and depositors alike. While data set content may not be included in this open access argument, there are other potential conflicts specifically related to research data. For example, there is a conflict that emerges in more local contexts related to defining the boundaries between IRs and university archives: Who gets to "own" this research content? However, it is likely that increasing institutional demands for stewardship of research data will require the attention and resources of both the library and the archives. For some institutions, resultant decisions on stewardship of research data and long-lived digital data collections (National Science Board, 2005) will foster new relationships between these often separate organizations, and may necessitate negotiation of the treatment of the university's scientific record.

If the purpose and content of an IR varies from institution to institution, what are the common characteristics of institutional repositories that we can identify?

The content in IRs tends to be freely available to anyone with access to the Internet, although there are sometimes access restrictions or embargos placed on material. In order to promote long-term access, most IRs offer persistent URLs—that is, Web addresses that do not change—using Handles (http://www.handle.net/), Archival Resource Keys (ARK) (http://www.cdlib.org/inside/diglib/ark/), or other services. Institutional repositories also are likely to be optimized for crawling by Web spiders from search engines like Google; "splash" pages of the information describing material (metadata) are crawled as well as the content itself. Most IRs expose their metadata for harvesting via the Open Archives Initiative Protocol for Metadata Harvesting (OAI PMH) (http://www.openarchives.org/). The OAI PMH facilitates an integrated search with content from other institutional repositories (as well as any other service exposing metadata via the OAI PMH) via a service like OAIster (http://www.oaister.org/). Many offer RSS feeds and other services that further the dissemination of the material held in the repository. All of these efforts serve the purpose of maximizing the discovery of and access to the content contained in the repository, as well as the promotion of the institution as a whole.

IRs also try to collect content and some amount of descriptive information either through direct deposit by the author(s) or by an intermediary (including librarians). The depositor is expected to have the right to deposit or to have negotiated the right to deposit the content, although we have found that in practice it is often the repository managers who are doing this work. Typically, materials in IRs do not go through the normal library or archives acquisitions routes; and unlike traditional library materials, they usually do not receive full cataloging. However, IRs do generally have policies in place that define who can deposit, what may be deposited, what metadata is required, etc. (Markey, Rieh, St. Jean, Kim, & Yakel, 2007, pp. 46–47). IRs typically do not have any formal peer review or editorial process for deposited items; the researcher's institutional or organizational affiliation is usually considered the de facto authentication for deposit. In addition, many, but not all, repositories make some level of commitment to the long-term preservation of and persistent access to the material contained within them.

IRs now exist in a rather contested space. Early hopes that IRs would impact the scholarly publishing sphere have been dampened by low self-archiving rates (McDowell, 2007). As noted above, the collection and preservation of grey literature, data sets, and other unpublished material has been called a distraction from self-archiving and open access, touted as the "primary" purpose of IRs by some open access advocates. Current software packages often are inadequate to manage the range of materials collected (particularly data sets) and fail to provide the range of services that repository managers are finding their users want. Preservation—a stated goal of many IRs—is often an afterthought, as repository managers focus on getting material into IRs first (see Rieh et al., in this issue). In many cases, IRs are not part of a larger vision for services the library can provide to the institution, but are isolated projects without a strong base of support (see Salo and Rieh et al. in this issue). Given all of these factors, where will IRs be in the next five or ten years?

This issue of *Library Trends* contains an impressive slate of articles from prominent practitioners and researchers in the field, who offer a range of perspectives on the current state of IRs in academic institutions and reflections on their future. We start with Dorothea Salo's "The Innkeeper at the Roach Motel." Salo, an experienced repository manager, unflinchingly describes the current dilemma of institutional repositories, that they are predicated on unrealistic goals. Salo takes apart much of the hopeful rhetoric that surrounds institutional repositories and in its place offers a realistic way forward.

The United Kingdom has been notable in their efforts to fund a range of implementation and research projects around IRs, particularly through funding from the Joint Information Systems Committee (JISC). In "Institutional Repositories in the UK: the JISC Approach," Neil Jacobs, Amber

Thomas, and Andrew McGregor of JISC describe the rationale behind this funding and the impressive results it has brought about.

We move then to articles describing research results from two important research projects on IRs: the Mellon funded "Identifying Factors of Success in CIC Institutional Repository Development" project led by Carole Palmer at the University of Illinois at Urbana-Champaign and the MIRACLE Project led by Soo Young Rieh at the University of Michigan. In "Strategies for Institutional Repository Development: A Case Study of Three Evolving Initiatives," Palmer, Teffeau, and Newton analyze the choices made during the development of three university IRs, with the aim of distinguishing the conditions that are driving IR development. They discuss three different approaches to IR development, and describe how decisions on IR implementation are based on "balancing content acquisition and service provision" within the complicated (and varied) landscape of library management and resource distribution, emergent roles for library liaisons, technical infrastructure, and intellectual property issues. Rieh, St. Jean, Yakel, Markey, and Kim describe the results of their telephone interviews with staff involved in IRs in "Perceptions and Experiences of Staff in the Planning and Implementation of Institutional Repositories." Although staff remain enthusiastic about the development and implementation of IRs (and interestingly, in light of Salo's essay, in their role in the open access movement), Rieh and her coauthors do find that IRs have yet to be implemented as part of a coherent set of services.

Three of the papers in this issue (introduced below) specifically address data curation and IR activities at research universities. Data curation is the active and ongoing management of data through its lifecycle of interest and usefulness to scholarship, science, and education, which includes appraisal and selection, representation and organization of these data for access and use over time. "[C]uration embraces and goes beyond that of enhanced present-day re-use, and of archival responsibility, to embrace stewardship that adds value through the provision of context and linkage: placing emphasis on publishing data in ways that ease re-use and promoting accountability and integration" (Rusbridge et al, 2005, p. 2). Adding to library collections raw or processed research data sets, that is, data forms other than final published results or derivative secondary resources, brings new complexities to the conduct of traditional library processes. Library and archival professionals are beginning to consider these issues that are facing the field. Davis and Vickery (2007) identify new challenges specific to collection development practices that will come with adding data sets. They suggest that budgetary constraints may be restricting the adoption of rational appraisal and selection policies, and that future planning will need to account for "hidden costs associated with collecting data sets" (p. 30). Green and Gutmann (2007) discuss the necessity for collaborations and cooperative arrangements among IRs, disciplinary and community repositories, and large data centers.

As Anderson (2004) stated, "Archiving and preservation of S&T [science and technology] data can no longer be thought of as a post project activity" (p. 191). We are witnessing already the pressures of this requirement: there are new divisions of labor emerging in the academic environment to address the needs and requirements for the management and preservation of research data, and the lack of a sufficient infrastructure and a trained workforce is evident (Association of Research Libraries, 2007; Swan & Brown, 2008).

As a result of these needs and changes, a variety of models are emerging to address the acquisition and stewardship of scientific research data by academic libraries across the United States. These models span a range of approaches, from consultation on data management and curation services, to direct collaborations with scientists, where library services are added directly into grant proposals and researchers work with library staff to develop tools together. For some libraries, there is active participation at the institutional level to mobilize for e-Research and e-Learning. Three of the papers in this issue specifically address data curation and related IR activities at research universities. The authors present the decision-making context and evolving data stewardship trajectory unique to their individual institutions. A goal for each of these libraries is to provide an organizational structure and technical infrastructure that will facilitate data-intensive science and scholarship.

For the Purdue University Libraries, the architectural approach to developing services for, and access to, university content is to build a distributed IR consisting of three parallel repositories for electronics documents, the digital archives, and research data. In his article, "Institutional Repositories and Research Data Curation in a Distributed Environment," Michael Witt makes a case for understanding the (traditional) scientific process, and its role in shaping their thinking about the organization of data services. He goes on to explain some of the fundamental technical requirements necessary for providing data curation services through a library-based IR model.

In "At the Watershed: Preparing for Research Data Management and Stewardship at the University of Minnesota Libraries," Leslie Delserone considers the convergence of several library administration components—hiring, research, and program development—and their impact on moving data stewardship activities forward at the University of Minnesota's Libraries. Delserone presents these changes in the library within the broader context of a systematic engagement with e-Science activities at the university level, and collaborative efforts with UMN Office of Information Technology, the UMN IR, and the university's Research Cyberinfrastructure Alliance.

In "Case Study in Data Curation at Johns Hopkins University" Sayeed Choudhury offers a reasoned view of repositories that acknowledges their simultaneous promise but, thus far, lack of anticipated success. At the

Johns Hopkins University Libraries, the view is that the IR and data curation infrastructure are components of services that need to be developed in response to their researchers' scientific work process and scholarly communication. Choudhury points out that, in order to serve the needs of scientists and scholars, "it will be necessary to move away from a collection or institution-centric view . . . to serve scholars most effectively." They are working, therefore, to integrate the library IR into "a larger landscape of repositories."

There are several similar aspects to these three cases that are noteworthy: They share a common goal to provide an organizational structure and infrastructure that will facilitate data-intensive science and scholarship at their respective universities. All three of the libraries represented here are creating new ways for librarians to engage with scientists and to participate in data curation activities. Also common to these models is the development of new positions that expand librarians' roles, thus adding organizational capacity to address growing demand for data curation. This is evidenced by new job titles we are seeing across the field, including "Data Humanist," "Data Scientist," "Interdisciplinary Research Librarian," "Data Research Scientist," and "Data Services Librarian."[3]

In the latter part of this issue we have four articles that look at the future of IRs.

Julie Allinson in "Describing Scholarly Works with Dublin Core: A Functional Approach" describes important work to create an application profile to describe materials appropriate for institutional repositories. Allinson outlines use cases that rely on metadata that is far more expressive and flexible than what is currently available in most IR software packages and describes the process of developing an application profile that fulfills these requirements. We chose to place this piece with other articles about the future of IRs, for while the Scholarly Works Application Profile (SWAP) is an established profile, it has yet to be fully implemented in IR software systems.

Ellen Finnie Duranceau, the Scholarly Publishing and Licensing Consultant for the MIT Libraries, describes the future of IRs within the context of Yochai Benkler's *The Wealth of Networks* in "The 'Wealth of Networks' and Institutional Repositories: MIT, DSpace, and the Future of the Scholarly Commons." Using MIT as an example, Duranceau moves forward from Salo's work to describe how libraries could conceive of institutional repositories as a piece of a larger movement towards openness on campus.

Nancy McGovern and Aprille McKay focus on IRs and preservation in their piece "Leveraging Short-term Opportunities to Address Long-term Obligations: A Perspective on Institutional Repositories and Digital Preservation Programs." McGovern and McKay's experience and their interactions with many institutions in the area of digital preservation give them valuable perspective on both the pitfalls and potentials for IRs in the realm

of digital preservation. It is instructive to read this article after Rieh et al.'s finding that preservation is often an afterthought for IR staff because of pressures to fill the repository.

Finally, Bryan Heidorn's article titled, "Shedding Light on the Dark Data in the Long Tail of Science," concerns discovery and access for "dark data in the long tail of science," that is, data which are hidden from would-be users. These data have high potential value but are at great risk for loss, which can result in unnecessary replication and missed opportunities for new scientific discovery. Heidorn provides a detailed analysis of the characteristics of data in the "long tail," and then presents a set of possible solutions to easing current barriers to acquiring and preserving these data.

Institutional repositories currently exist in a rapidly shifting landscape, and there appears to be no definite consensus on what their role might be in the future. However, it is possible to see glimpses of what is to come as one reads through the range of perspectives in this issue: the IR as infrastructure to support data curation and long-term preservation of digital materials produced on campus; the IR as one of a set of services to support scholarly communication and openness on campus; the IR as a tool for librarian subject liaisons to work more closely with their faculty and move closer to the research process; the IR as a collaborative activity between libraries, university archives, and campus IT organizations. We hope that this issue will further discussion and move the community toward a better understanding of institutional repositories and their place in academic institutions.

Notes

1. Two useful resources for exploring the range of repositories are the Directory of Open Access Repositories (OpenDOAR) at http://www.opendoar.org/ and the Registry of Open Access Registries (ROAR) http://roar.eprints.org/.
2. We use the term data sets to mean research data generated from science, the social sciences, or the humanities.
3. While this job title is not new to social science librarians, the position scope for a data services librarian in some libraries is moving well beyond traditional services supporting the use of social science data.

References

Association of Research Libraries, Joint Task Force on Library Support for E-Science. (2007). Agenda for developing E-science in research libraries. Washington, D.C.: Association of Research Libraries. Retrieved October 14, 2008, from http://www.arl.org/bm~doc/ARL_EScience_final.pdf

Anderson, W. L. (2004). Some challenges and issues in managing, and preserving access to, long-lived collections of digital scientific and technical data. *Data Science Journal, 3*(30), 191–202.

Crow, R. (2002). *The Case for Institutional Repositories: A SPARC Position Paper.* Washington, D.C.: SPARC. Retrieved April 28, 2008, from http://www.arl.org/sparc/bm%7Edoc/ir_final_release_102.pdf

Davis, H., & Vickery, J. (2007). Datasets, a shift in the currency of scholarly communication: Implications for Library collections and acquisitions. *Serials Review, 33*(1), 26–32.

Green, A. G., & Gutmann, M. P. (2007). Building partnerships among social science researchers, institution-based repositories, and domain specific data archives. OCLC Systems and Services. *International Digital Library Perspectives, 23*(1), 35–53.

Harnad, S., Brody, T., Valliéres, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H., & Hilf, E.R. (2004). The access/impact problem and the green and gold roads to open access. *Serials Review, 30*(4), 310–314.

Lynch, C.A. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. *ARL: A Bimonthly Report* no. 226. Retrieved April 28, 2008, from http://www.arl.org/resources/pubs/br/br226/br226ir.shtml

Markey, K., Rieh, S. Y., St. Jean, B., Kim, J., & Yakel, E. (2007). Census of Institutional Repositories in the United States: MIRACLE Project Research Findings. Washington, D.C.: CLIR. Retrieved February 26, 2007, from http://www.clir.org/pubs/reports/pub140/pub140.pdf

McDowell, C. S. (2007). Evaluating institutional repository deployment in American academe since early 2005: Repositories by the numbers, Part 2. *D-Lib Magazine, 13* (9/10). Retrieved April 28, 2008, from http://www.dlib.org/dlib/september07/mcdowell/09mcdowell.html

National Science Board. (2005). Long-Lived Digital Data Collections: Enabling research and education in the 21st Century. NSB-05-40. Available: http://www.nsf.gov/pubs/2005/nsb0540/

Poynder, R. (2006). Clear blue water. Retrieved May 7, 2008, from http://poynder.blogspot.com/2006/03/institutional-repositories-and-little.html

Prosser, D. C. (2003). Scholarly communication in the 21st century—The impact of new technologies and models. *Serials, 16*(2), 163–167.

Rusbridge, C., Burnhill, P., Ross, S., Buneman, P., Giaretta, D., Lyon, L., & Atkinson, M. (2005). The Digital Curation Centre: A vision for digital curation. In proceedings of Local to Global Data Interoperability—Challenges and Technologies, 2005. Mass Storage and Systems Technology Committee of the IEEE Computer Society, 20–24 June 2005, Sardinia, Italy. Retrieved October 29, 2008, from http://eprints.erpanet.org/82/

Swan, A., & Brown, S. (2008). The skills, role and career structure of data scientists and curators: An assessment of current practice and future needs. A report to the Joint Information Systems Committee (JISC), July, 2008. Retrieved October 14, 2008, from http://www.jisc.ac.uk/media/documents/programmes/digitalrepositories/dataskillscareersfinalreport.pdf

Thomas, C., & MacDonald, R. (2007). Measuring and comparing participation patterns in digital repositories: Repositories by the numbers, Part 1. *D-Lib Magazine, 13*(9/10). Retrieved October 10, 2008, from http://www.dlib.org/dlib/september07/mcdonald/09mcdonald.html

---

Sarah L. Shreeves is the coordinator for the Illinois Digital Environment for Access to Learning and Scholarship (IDEALS), the set of services and collections that comprise the institutional repository at the University of Illinois at Urbana-Champaign (UIUC). In her position, Sarah is responsible for working with faculty, students, and researchers on a range of scholarly communication issues including author rights, open access, and data curation. She has spoken and published on metadata interoperability particularly within the context of the Open Archives Initiative Protocol for Metadata Harvesting. Her research interests lie in scholarly communication issues related to institutional repositories and metadata interoperability. She has a BA in Medieval Studies from Bryn Mawr College, an MA in Children's Literature from Simmons College, and an MS in Library and Information Science from UIUC.

Melissa H. Cragin is project coordinator for the Data Curation Education Program (DCEP) at the Graduate School of Library and Information Science (GSLIS) at the University of Illinois at Urbana-Champaign. She is a coprincipal investigator on the IMLS funded project, "Investigating Data Curation Profiles Across Multiple Research Disciplines," which is being led by the Purdue University Libraries. Melissa is also a doctoral candidate at GSLIS, where her research concerns emergent changes in formal and informal scholarly communication, with particular emphasis on scientific work practices related to data.