ARL

A BIMONTHLY REPORT ON RESEARCH LIBRARY ISSUES AND ACTIONS FROM ARL, CNI, AND SPARC

226 February 2003

Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age

by Clifford A. Lynch, Executive Director, Coalition for Networked Information

Introduction

In the fall of 2002, something extraordinary occurred in the continuing networked information revolution, shifting the dynamic among individually driven innovation, institutional progress, and the evolution of disciplinary scholarly practices. The development of institutional repositories emerged as a new strategy that allows universities to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication, both moving beyond their historic relatively passive role of supporting established publishers in modernizing scholarly publishing through the licensing of digital content, and also scaling up beyond ad-hoc alliances, partnerships, and support arrangements with a few select faculty pioneers exploring more transformative new uses of the digital medium.

Many technology trends and development efforts came together to make this strategy possible. Online storage costs have dropped significantly; repositories are now affordable. Standards like the open archives metadata harvesting protocol are now in place; some progress has also been made on the standards for the underlying metadata itself. The thinking about digital preservation over the past five years has advanced to the point where the needs are widely recognized and well defined, the technical approaches at least superficially mapped out, and the need for action is now clear. The development of free, publicly accessible journal article collections in disciplines such as high-energy physics has demonstrated ways in which the network can change scholarly communication by altering dissemination and access patterns; separately, the development of a series of extraordinary digital works had at least suggested the potential of creative authorship specifically for the digital medium to transform the presentation and transmission of scholarship.

The leadership of the Massachusetts Institute of Technology (MIT) in the development and deployment of the DSpace institutional repository system <http://www.dspace.org/>, created in collaboration with the Hewlett Packard Corporation, has been a model pointing the way forward for many other universities. In 2003, with funding from The Andrew W. Mellon Foundation and other sources, MIT's DSpace is scheduled to be replicated at a number of additional institutions around the world; the software has also been released publicly under an open source arrangement, greatly lowering the cost and development barriers to implementing repositories for all institutions. The MIT software is not the only option available, although I believe it is the most general-purpose; for example, there is software from the University of Southampton in the U.K. <http:// www.eprints.org/> designed more specifically for institutional or disciplinary repositories of papers, as opposed to arbitrary digital materials.

Over the past few months, I have had a number of opportunities to speak about the roles and significance of institutional repositories as a strategy for supporting the use of networked information to advance scholarship, notably at a workshop jointly sponsored by ARL, CNI, and SPARC in Washington, D.C., at the DSpace launch celebration at MIT, and at the University of Tennessee and the University of British Columbia. While video recordings of some of these events are available on

the Net, this article is an attempt to summarize and articulate the views I've expressed at these various events about the nature and functions of institutional repositories and their role in transforming scholarship.

Defining Institutional Repositories

In my view, a university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution. While operational responsibility for these services may reasonably be situated in different organizational units at different universities, an effective institutional repository of necessity represents a collaboration among librarians, information technologists, archives and records mangers, faculty, and university administrators and policymakers. At any given point in time, an institutional repository will be supported by a set of information technologies, but a key part of the services that comprise an institutional repository is the management of technological changes, and the migration of digital content from one set of technologies to the next as part of the organizational commitment to providing repository services. An institutional repository is not simply a fixed set of software and hardware.

a key part of the services that comprise an institutional repository is the management of technological changes, and the migration of digital content from one set of technologies to the next as part of the organizational commitment to providing repository services.

While early implementers of institutional repositories have chosen different paths to begin populating their repositories and to build campus community acceptance, support, and participation, I believe that a mature and fully realized institutional repository will contain the intellectual works of faculty and students—both research and teaching materials—and also documentation of the activities of the institution itself in the form of records of events and performance and of the ongoing intellectual life of the institution. It will also house experimental and observational data captured by members of the institution that support their scholarly activities.

At the most basic and fundamental level, an institutional repository is a recognition that the intellectual life and scholarship of our universities will increasingly be represented, documented, and shared in digital form, and that a primary responsibility of our universities is to exercise stewardship over these riches: both to make them available and to preserve them. An institutional repository is the means by which our universities will address this responsibility both to the members of their communities and to the public. It is a new channel for structuring the university's contribution to the broader world, and as such invites policy and cultural reassessment of this relationship.

I want to make the distinction between scholarly publishing as it is currently practiced and the broader, much more diverse, often less formal, and certainly more rapidly evolving set of practices that comprise scholarly communication; scholarly publishing is a very specific, circumscribed example of scholarly communication. I use the two terms "scholarly communication" and "scholarly publishing" distinctly and carefully in this paper. For example, the definition I propose for an institutional repository does not call for a new scholarly publishing role for universities, only one of dissemination of scholarly communication; scholarly publishing is much more than simple dissemination, and has typically been rather limited in the genres of communication that it does disseminate. I will have more to say about the relationships between repositories and publishing later.

For those organizations *within* the university concerned with stewardship—we think immediately of libraries, archives, and museums but should recognize there are also huge numbers of academic units that curate collections of information—it should be clear that institutional repositories raise complex and nuanced questions about organizational roles, responsibilities resources, and strategies. Similar, but perhaps less complex, questions arise for all organizational units focused on dissemination of scholarly communication or more narrowly on scholarly publishing, such as university presses.

The Strategic Importance of Institutional Repositories

Scholarship and scholarly communication are changing. These changes start with risky and bold acts of individual creativity. They will extend slowly to cultural changes at the disciplinary level and ultimately to new interdisciplinary standards that are expressed in the decisions of institutional tenure and promotion practices.

Our institutions of higher education have overlooked an opportunity to support our most innovative and creative faculty for at least a decade now, to the detriment of both the faculty members and the institutions themselves. These faculty have been

exploring ways in which works of authorship in the new digital medium can enhance teaching and learning and the communication of scholarship; such innovations are essential to keeping scholarship vital and effective, and they must not only be supported but nurtured. Indeed nurturing these innovations reaches to the core mission of our universities, and to the core values of our universities. A much broader and generally more conservative group of faculty have exploited the Net as a vehicle for sharing their ideas worldwide, whether these ideas are expressed in relatively familiar forms such as digital versions of traditional journal articles or (less commonly) in entirely new forms that begin to map out the future evolution of, for example, the scholarly monograph in the digital medium. This embrace of new dissemination opportunities is also important for what it says about the roles of scholars and universities in society and in a global environment. Our universities have poorly served this broader group of scholars as well, though this may be less critical because faculty are well motivated to rise above the institutional failures to help them disseminate their works, because failures to effectively disseminate these works are less damaging than failures to legitimize nontraditional works, and because faculty concerned only with dissemination of traditional material are at less risk within their own disciplines.

an institutional repository is a recognition that the intellectual life and scholarship of our universities will increasingly be represented, documented, and shared in digital form, and that a primary responsibility of our universities is to exercise stewardship over these riches: both to make them available and to preserve them.

But consider the plight of a faculty member seeking only broader dissemination and availability of his or her traditional journal articles, book chapters, or perhaps even monographs through use of the network, working in parallel with the traditional scholarly publishing system. Such a faculty member faces several time-consuming problems. He or she must exercise stewardship over the actual content and its metadata: migrating the content to new formats as they evolve over time, creating metadata describing the content, and ensuring the metadata is available in the appropriate schemas and formats and through appropriate protocol interfaces such as open archives metadata harvesting. Faculty are typically best at creating new knowledge, not maintaining the record of this process of creation. Worse still, this faculty member must not only manage content but must manage a dissemination system such as a personal Web site, playing the role of system administrator (or the manager of someone serving as a system administrator). Over the past few years, this has ceased to be a reasonable activity for most amateurs; software complexity, security risks, backup requirements, and other problems have generally relegated effective operation of Web sites to professionals who can exploit economies of scale, and who can begin each day with a review of recently issued security patches. Today, our faculty time is being wasted, and expended ineffectively, on system administration activities and content curation. And, because system administration is ineffective, it places our institutions at risk: because faculty are generally not capable of responding to the endless series of security exposures and patches, our university networks are riddled with vulnerable faculty machines intended to serve as points of distribution for scholarly works. And faculty create content at risk because they typically do not back it up appropriately, ensure its integrity (in part by hosting it on secure systems), and curate it properly.

For those faculty who are concerned not just with distribution opportunities through the network but with deeper questions of how to exploit the nature of the digital medium for new works of authorship, the situation is even worse. This is not just about more effective public access to recognizable and familiar genres of work such as journal articles which can, in the worst case, be reduced to printed forms for distribution to a tenure and promotion committee. These faculty take on a heavy burden in arguing for the legitimacy of investing their time in works of digital scholarship, and in making the case for the value of such creations in comparison to more traditional scholarly output. This is a cultural problem that must be played out discipline by discipline, and which must be worked out also in the evaluation, tenure, and promotion practices in place at an institutional level. However, preservability is an essential prerequisite to any claims to scholarly legitimacy for authoring in the new medium; without being able to claim such works are a permanent part of the scholarly record, it's very hard to argue that they not only deserve but demand full consideration as contributions to scholarship. Most individual faculty lack the time, resources, or expertise to ensure preservation of their own scholarly work even in the short term, and clearly can't do it in the long term that extends beyond their careers; the long term can *only* be addressed by an organizationally based strategy. Institutional repositories

can address both the near-term questions about continuity of access by providing an environment in which such new works of scholarship can be managed and disseminated—including such basic things as professionally managed systems and systematic backup procedures—and also the longer-term questions about preservation by creating an institutional commitment to such preservation.

The revolution in scholarly communications is not limited to the development of new genres of scholarly works that are enabled by the digital medium; even traditional forms such as journal articles now frequently include supplementary datasets and analysis tools. Scholarship has become data intensive; it is supported and documented by data and tools that complement interpretive works of authorship. For the sciences, these changes have been well documented in the recent National Science Foundation report of the Advisory Committee for Cyberinfrastructure chaired by Dan Atkins;[1] while the report is focused on cyberinfrastructure to support the conduct of science, most of the discussion is in fact applicable beyond the sciences to the broader scholarly enterprise, including the humanities. Most scientific journals are now accepting what they characterize as "supplementary" materials as part of the publication of traditional journal articles, but it is much less clear what commitments these journals are making to actually integrating these supplementary materials into the permanent record of scholarship in the same way that they maintain the journal articles themselves as a part of that record. While it is clear that for some types of scholarly work we will see the continued evolution of disciplinary data repositories (consider, for example, molecular biology) and community norms that journal publication is complemented by deposit of data in these disciplinary repositories, it is equally clear that the scholarly enterprise is sufficiently diverse that these disciplinary repositories will never be fully comprehensive. Only an institutionally based approach to managing these data resources, which operates in alignment with what the faculty at each individual institution are actually doing, can provide a comprehensive dissemination and preservation mechanism for the data that supports the new scholarship for the digital world. Journals will move too slowly and too unevenly to manage these resources, and disciplinary data repositories cannot be comprehensive. Institutional repositories can maintain data in addition to authored scholarly works. In this sense, the institutional repository is a complement and a supplement, rather than a substitute, for traditional scholarly publication venues.

Institutional repositories also have roles beyond disseminating and managing the works of individual scholars that are part of the dialog of scholarly communications. I have argued that research libraries must establish new collection development strategies for the digital world, taking stewardship responsibility for content that will be of future scholarly importance. Institutional repositories are a place where they can put much of the material that research libraries identify as worth collecting. Finally, at least a few institutions themselves are changing their culture and are making commitments to globally disseminate extensive teaching and learning materials through the Net (for example, the OpenCourseWare initiative at MIT <http://ocw.mit.edu/>), or, at a less systematic but still important level, to digitally capture and preserve the many of the events of campus life—symposia, performances, lectures. Institutional repositories offer a framework for organized stewardship and accessibility of these materials.

To summarize, institutional repositories can facilitate greatly enhanced access to traditional scholarly content by empowering faculty to effectively use the new dissemination capabilities offered by the network. This is also occurring on a disciplinary basis through the development of e-print and preprint servers, at least in some disciplines. In cases where the disciplinary practice is ready, institutional repositories can feed disciplinary repositories directly. In cases where the disciplinary culture is more conservative, where scholarly societies or key journals choose to hold back change, institutional repositories can help individual faculty take the lead in initiating shifts in disciplinary practice.

Institutional repositories can encourage the exploration and adoption of new forms of scholarly communication that exploit the digital medium in fundamental ways. This, to me, is perhaps the most important and exciting payoff: facilitating change not so much in the existing system of scholarly publishing but by opening up entire new forms of scholarly communication that will need to be legitimized and nurtured with guarantees of both short- and long-term accessibility. Institutional repositories can support new practices of scholarship that emphasize data as an integral part of the record and discourse of scholarship. They can structure and make effective otherwise diffuse efforts to capture and disseminate learning and teaching materials, symposia and performances, and related documentation of the intellectual life of universities.

Cautions about Institutional Repositories

There are at least three areas in which I am concerned attempts to develop institutional repositories could go seriously astray and become counterproductive.

The first potential danger is that institutional repositories are cast as tools of institutional (administrative) strategies to exercise control over what has typically been faculty controlled intellectual work.

I believe that any institutional repository approach that *requires* deposit of faculty or student works and/or uses the institutional repository as a means of asserting control or ownership over these works will likely fail, and probably deserves to fail. Institutional repositories will succeed precisely because they are responsive to the needs of campus communities, and advance the interests of campus communities and of scholarship broadly. To the extent that they try to *enforce* behavioral or cultural changes—and particularly controversial ones—within the campus community they will and should fail. The theme is accepting responsibility, not exerting new levels of control. This is not to say that policies mandating the deposit of materials that are broadly recognized as part of the institutional record (and recognized as being owned by the institution itself) are inappropriate. But institutions should move very conservatively down this path.

My second concern is somewhat similar to the first, that we respect institutional repositories as infrastructure and not overload this infrastructure with distracting and irrelevant policy baggage, but from a very different perspective.

We must not lose the crucial distinction between the role of institutions in establishing institutional repositories and the roles of scholarly communities within the institution's organizational units or within disciplines in creating and managing scholarly communication mechanisms that may build upon an institutional repository infrastructure. Campus administrators, librarians, and faculty members wishing to challenge existing systems of scholarly publishing (specifically their economic models and their creation of barriers to access through intellectual property control and licensing arrangements) may try to link their efforts too directly to institutional repositories by imposing inappropriate policy constraints upon the repository services.

Institutional repositories may legitimately serve as *infrastructure* to advance some of these interests—for example, groups might construct a peer-review process that certifies selected works that are accessible in various institutional repositories and even develop overlay systems that span a complex of institutional repositories and create a "virtual" journal. Note that such an effort would have to be extra-institutional and cross-institutional to have much scholarly credibility, for the same reason that university presses aren't simply publishing outlets for the faculty at their parent institutions, and the editorial boards of institutionally hosted journals are drawn from beyond the host institution. Its extra-institutional nature should help to clarify that it shouldn't be confused with the development of individual institutional repositories.

But this is not, to my mind, the primary point of institutional repositories. Indeed, it dramatically underestimates the importance of institutional repositories to characterize them as instruments for restructuring the current economics of scholarly publishing rather than as vehicles to advance, support, and legitimize a much broader spectrum of new scholarly communications. Further, I would argue that complex, cumbersome "gate keeping" policies for admitting materials to institutional repositories—particularly those that emulate practices from traditional scholarly publication such as the use of peer reviewers—are highly counterproductive; this will prevent institutional repositories from supporting and empowering faculty innovators and leaders. Membership in the campus community—certainly, if nothing else, membership in the campus faculty—should be sufficient credential to place materials in the institutional repository. To be sure, there are practical resource constraints that each institution will have to work out; some faculty have truly enormous datasets or multimedia collections that may be hard to accommodate. But recognize that the institutional repository isn't a journal, or a collection of journals, and should not be managed like one. That's not the point or the purpose of an institutional repository.

We must not lose the crucial distinction between the role of institutions in establishing institutional repositories and the roles of scholarly communities within the institution's organizational units or within disciplines in creating and managing scholarly communication mechanisms that may build upon an institutional repository infrastructure.

This does not preclude erecting superstructures on top of an institutional repository that implement elaborate gate-keeping mechanisms (the "community" mechanisms in DSpace, for example, allow the devolution of policies to specific groups and also sub-branding of areas within the repository as being under the policy control of specific groups) but the key point is that the basic repository service is an infrastructure service that should be kept divorced from policies imposed by such overlays. Such overlays might represent new journals, as already discussed; they might also represent archives, complete with appraisal systems and record-retention schedules, for example. My

argument is simply that it's important to maintain a simple, low-barrier-to-submission, basic repository service as well, and that this service is much of the point of setting up the repository in the first place.

Institutional repositories are not a challenge or alternative to disciplinary repositories; rather, they complement them, just as they can complement existing venues of scholarly publication. The Open Archive Initiative Metadata Harvesting Protocol <http://www.openarchives.org/> gives us the tools for an institutional repository to act as an entry point for redistributing works to systems of disciplinary repositories. It is desirable to make this as simple as possible, and to insulate faculty from having to deal with the details of a constantly evolving multiplicity of disciplinary services. Better to present the faculty with a simple and stable submission interface to the institutional repository. In this sense institutional repositories can be an infrastructure upon which disciplinary services and repositories can build.

I have a third, rather different, concern about institutional repositories. We are now seeing a substantial number of leading institutions making commitments to implement them. In the near future, many campus communities may expect and demand that such services be made available rapidly; creating institutional repositories may also become fashionable in some administrative circles. My fear is that, at some institutions, repositories will be offered hastily and without much real institutional commitment.

It's vital that institutions recognize institutional repositories as a serious and long-lasting commitment to the campus community (and to the scholarly world, and the public at large) that should not be made lightly. In establishing institutional repositories, institutions are both accepting risks and making promises; they are creating new expectations. In a budget crunch, the institutional repository may be one of the last things that can be cut, given the way that digital preservation demands steady and consistent attention and hence funding. Faculty who choose to rely on institutional repositories to disseminate and preserve their work are placing a great deal of trust in their institution and in the integrity, wisdom, and competence of the people who manage it. We need to ensure that our institutional repositories are worthy of this trust.

An institutional repository can fail over time for many reasons: policy (for example, the institution chooses to stop funding it), management failure or incompetence, or technical problems. Any of these failures can result in the disruption of access, or worse, total and permanent loss of material stored in the institutional repository. As we think about institutional repositories today, there is much less redundancy than we have had in our systems of print publication and libraries, so any single institutional failure can cause more damage. I worry a great deal about what the various impacts and implications of the first few major failures of institutional repositories—for whatever reasons—will be; I fear, for example, that they may greatly set back scholarly acceptance of authorship of digital works; they may have a corrosive effect on the trust that underpins campus communities; they may undermine broad social support for higher education. Sadly, I have little doubt that we *will* see such failures within the next decade or so. I hope I am wrong.

> *Well designed institutional repositories will separate system operation from curatorial and policy control (e.g. submission, preservation, etc) of specific sets of content. Thus we can expect institutional repositories to be a basic part of the negotiations in the development of regional or disciplinary consortia among universities or libraries.*

Stewardship is easy and inexpensive to claim; it is expensive and difficult to honor, and perhaps it will prove to be all too easy to later abdicate. Institutions need to think seriously before launching institutional repository programs.

Institutional Repositories and Networked Information Standards and Infrastructure

I believe that institutional repositories will promote progress in the development and deployment of infrastructure standards in a variety of difficult or neglected areas. Here I'll mention only three.

Preservable Formats. Institutional repositories make promises about stewardship and preservation. These promises are necessarily qualified. Institutions will make choices based on a balance of campus community demand and local assessments about technical feasibility, which will result in lists of file formats that they will commit to preserve in accessible forms (presumably through format migration); in other cases, they may preserve the bits that make up a file, but will offer no guarantees that these bits can be interpreted in the future without the development of specialized programs to read them. These choices can then be collected into broader community consensus within the higher education and research community as a form of bottom-up standards development that benefits from active work in curation and ongoing faculty involvement.

Identifiers. The ability to make persistent reference to materials in institutional repositories will clearly be

critical as these materials will form an important part of the scholarly dialogue and record. This will have to include provisions to deal with issues like versioning. Higher education and the library community have not been sufficiently active in this area, largely ceding the field to commercial interests and traditional publisher agendas. The deployment of institutional repositories will drive pragmatic solutions in this area.

Rights Documentation and Management. The management of rights for digital materials will be essential. The whole point of institutional repositories is to facilitate access, reuse, and stewardship (which may itself involve reformatting) of content, and we need methods of recording and documenting the rights and permissions associated with works that facilitate these goals of the research and education community. Part of this is a technical problem involving metadata structures; the other part is building consensus around a relatively small number of sets of terms and conditions that can cover the majority of the materials in practice. Working "standards" like the stock licenses under development by Creative Commons <http://creativecommons.org/> will be important here, and institutional repositories will be a way to make campus community members aware of these developments. Again, institutional repositories offer the opportunity for bottom-up, community-driven, consensus development about rights and permissions.

Future Developments in Institutional Repositories

I've described the current developments in institutional repositories and tried to explain why these are so deeply and strategically important to the enterprises of scholarship and higher education. The perspective has been largely a near-term one. In concluding this paper I want to at least sketch a few additional developments that may build upon an increasingly well established institutional repository model.

Not every higher education institution will need or want to run an institutional repository, though I think ultimately almost every such institution will want to offer some institutional repository services to its community. We will see various forms of consortial or cluster institutional repositories. Well designed institutional repositories will separate system operation from curatorial and policy control (e.g. submission, preservation, etc) of specific sets of content. Thus we can expect institutional repositories to be a basic part of the negotiations in the development of regional or disciplinary consortia among universities or libraries.

There is a clearly evolving idea of "federating" institutional repositories but as yet little concrete exploration of what this means—cross-repository search, swaps of storage between institutional repositories to gain geographic and systems diversity in pursuit of backup, preservation, and disaster recovery, or other capabilities. This will be a fruitful area for exploration and innovation. Another part of federation is that faculty often don't stay at a single institution for their entire career, and they frequently disregard institutional boundaries when collaborating with other scholars. Federation of institutional repositories may also subsume the development of arrangements that recognize and facilitate faculty mobility and cross-institutional collaborations.

Finally, university institutional repositories have some very interesting and unexplored extensions to what we might think of as community or public repositories; this may in fact be another case of a concept developed within higher education moving more broadly into our society. Public libraries might join forces with local government, local historical societies, local museums and archives, and members of their local communities to establish community repositories. Public broadcasting might also have a role here. In the long run it raises questions about "publishing" (and particularly nonprofit publishing) not in the scholarly context, but by members of arbitrary, perhaps but not necessarily geographically defined, communities or other interest groups. It is not inconceivable that we might also ultimately see commercial repository services for the public at large.

It is clear that the institutional repository is a very powerful idea that can serve as an engine of change for our institutions of higher education, and more broadly for the scholarly enterprises that they support. If properly developed, it advances a surprising number of goals, and addresses an impressive range of needs. Some of the results seem clear, though there are also likely to be any number of unexpected consequences. This is an area where I believe universities need to invest aggressively, but where they also need to implement thoughtfully and carefully, with broad consultation and collaboration across the campus community (with intellectual leadership from the faculty and the library working in partnership) and with a full understanding that if they succeed they will permanently change the landscape of scholarly communication.

[1] Atkins, Daniel E., et al., "Revolutionizing Science and Engineering through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure," January 2003, <http://www.communitytechnology.org/nsf_ci_report/>.