

INSTITUTIONAL REPOSITORIES

By Richard W. Boss

A number of academic libraries have created institutional repositories, digital collections capturing, preserving, organizing, and facilitating access to the intellectual output of a single college or university, or system of colleges and/or universities. The idea was not unique to them, but was borrowed from corporations that are research intensive and want to improve retention and access to the scholarly output of their researchers. While academic institutions generally seek to promote access to their repositories by those outside the institution, corporations generally seek to limit access to their own employees and the employees of companies with which they are engaged in collaborative research.

In January of 2006, the Technology Committee of the Public Library Association, the sponsor of TechNotes, decided that public libraries might have a role to play in creating and maintaining institutional repositories for the intellectual output of their communities. This TechNote investigates that question.

The Rationale

Underlying the concept of an institutional repository is the growing awareness that the traditional publishing model no longer meets the needs of those who seek to disseminate or access scholarly output. There is too much for the capacity of traditional publishing, both book and journal, to accommodate. The slowness and cost of disseminating scholarly output via print are also serious limitations. Increasingly, scholars are sharing their results with others via the Internet using e-mail, Web sites, blogs, etc. The problem is the lack of organization of the scholarly output, therefore, much that is worthwhile does not come to the attention of those who may find it useful; and much disappears after a short period of time.

The Elements of a Repository

There appears to be general agreement among those who have written about the subject that there are a few basic elements that characterize an institutional repository: an institution, scholarship, a digital collection, a retention format, and a retention period.

The first element of an institutional repository is an “*institution.*” Not only does that limit the scope of the collection, but it facilitates the establishment and maintenance of the repository because the institution presumably has an interest in promoting the scholarship of its faculty, students, or employees. It may authorize a library to assume this role and may even advance funds for it to do so.

A public library is itself an institution. In most cases, its staff will not produce enough scholarly output to warrant the establishment and maintenance of an institutional repository. A public library can think of itself as part of a greater whole, the community that it serves or the municipal or county government of which it is a part.

It is difficult to think of an entire community as an “institution” that a public library might seek to support with an institutional repository. Not only is a community comprised of multiple institutions, both academic and corporate, but also of many unaffiliated individuals. Clearly, a “community repository” established and maintained by a public library cannot expect to include all of the local scholarly output. For that reason, the role of the public library might be a combination of a repository and links to other repositories, thus facilitating the discovery of all publicly available scholarship, regardless of where the primary responsibility for capturing, preserving, organizing, and providing access lies. In that case, it would limit its collection to that which the other institutions do not include in their repositories.

Another option for a public library is to participate with other organizations, including

libraries, to jointly develop an institutional repository. This might significantly reduce the impact on staff and budget unless the consortium seeks to define the scope of the depository much more broadly than the library would have done.

A public library might also opt to define the institution as the municipal or county government of which it is a part. In that case, it would limit its collection to the output of that governmental entity. The question then arises as to whether it is acting as an electronic archive for governmental bodies and, if so, whether it has both a mandate and financial support for assuming that role.

The second essential element of a repository is a definition of “*scholarship*.” Not every piece of writing produced in the community, whether it is the entire service area or the government of which the library is a part is scholarly. A well reasoned paper that is footnoted might meet even narrowly drawn criteria, but what about a family history that has a great deal of information about the community? A major study of environmental issues by a government agency might also be considered scholarly. But the vast majority of public documents are merely intended to inform.

Perhaps a public library should focus not on scholarship, but on usefulness. If it will help a person understand the community or its government, it may be useful to capture, preserve, organize, and facilitate access to it even if it is not scholarly.

Many public libraries have collected published books and articles by local authors, books and articles about the community, local family histories, and publications of local government for many years. Perhaps, expanding the scope of the capture to include that which is in digital form should reflect the existing collection development policy’s scope, rather than being more broadly or narrowly defined.

The third element of an institutional repository is that it is a “*digital collection*.” That reflects the original rationale that traditional print media are not adequate for dissemination of all

information. Many of the articles on the subject published in the past five years assume that the materials have been created digitally, but should the repository be limited to that which is already in electronic form, or should it include images captured from printed material? If the latter, what should be scanned and who does the scanning? The priority might be on scanning that which is not commercially published and/or difficult to retain in print form. It is unlikely that a library will be able to avoid undertaking scanning when the creator of the material does not have the willingness or resources to do so.

The fourth element is the “*retention format.*” This element has not been given enough attention in the literature, especially when long-term retention is envisioned. The format must facilitate open access, rather than requiring the use of special software to retrieve and read the material. It must also not rely on a proprietary format that may be superseded. A proprietary data storage format depends on a single vendor remaining in business and continuing to offer fixes and upgrades at reasonable cost. WordPerfect and WordStar are now virtually extinct. Word appears to have a long life ahead of it, but a single company decides whether future versions will be compatible with older versions, and at what cost. There is a tendency for licensing costs to go up as new versions are released.

An open format that is not dependent on specific hardware or software that may become obsolete is essential. The most widely adopted “open format” standard is PDF ((Portable Document Format). It was originally a proprietary format, but Adobe, its developer, has released the PDF specifications to the public. While never formally adopted by any standards setting body, it has become a “de facto” standard. There are many companies that offer PDF products. Most documents being created today can be easily converted from their native file formats to PDF, and documents that are scanned from hard copies can be stored in PDF. No special reader is required; only a Web browser.

There is a competing format for paper documents that have been scanned, TIFF (Tagged Imaged File Format). The TIFF specifications are also owned by Adobe, but they have been released to the public so that anyone can use the specifications to develop TIFF products. There

is a drawback to TIFF: a TIFF viewer program is needed as the file cannot be seen in a browser or a word processing program the way that a PDF document can be.

There has been pressure on the International Organization for Standardization (ISO) to develop a standard based on PDF. It has begun work and released the first version of PDF/A (Portable Document Format/Archive). Designated ISO 19005.1, the standard is intended to be suitable for long-term preservation of page-oriented documents for which PDF is already being used. The major drawback is that a special viewer is required to read the file.

The fifth element is the *“retention period.”* The institutional repositories established and maintained by academic libraries have had as their goal the maintenance of the contents in perpetuity. That is based on the assumption that scholarship will always have at least historical value. A public library may choose to “weed” its digital collection just as it would a print collection. At a minimum, it may want to remove that which is subsequently published or made available in another repository to which its repository is linked. It may also be required to remove material that may subsequently be copyrighted by the author or were the copyright holder to terminate the license to disseminate the work.

The Impact on a Library

While it is not possible to project the cost of an institutional repository without knowing its scope and size, there is no doubt that capturing, maintaining, organizing and facilitating access will require staff time and money. The few academic institutional repositories with which the author is familiar have required hundreds to thousands of hours of staff time and have cost tens-of-thousands to hundreds-of-thousands of dollars. The University of Oregon has reported that its staff spends from 2,280 to 3,190 staff hours per year on its institutional repository. MIT is one of the few to quote its budget, \$285,000 per year.

Before pursuing the establishment of an institutional repository, a public library should undertake a needs assessment and determine what benefits will be realized by the community and at

what cost. Even given a positive cost benefit analysis, is this the best use of the available funds?

There is a good possibility that the first public libraries that establish institutional repositories will be able to obtain grants for doing so. Granting agencies are attracted to pioneering ventures, especially those that can act as models for others. Public libraries that have successfully completed grant-funded ventures are particularly favored by granting agencies. However, very few granting agencies are willing to fund operational costs once a venture has been launched. A public library should, therefore, estimate costs for a five-year period and assume that grant funds will be available for no more than the first two of those years.

Sources of Information

As of the second quarter of 2006, there appeared to be no sources in the published literature focusing on institutional repositories in public libraries. The best introduction to institutional repositories in academic libraries is a position paper by the Association of Research Libraries entitled

“The Case for Institutional Repositories: A SPARC Position Paper (www.arl.org/sparc/IR/ir.html/).

This source can be augmented by accessing the following selected institutional repositories:

- Boston College: <http://escholarship.bc.edu>
- MIT: <http://dspace.mit.edu/index/isp>
- University of California: <http://repositories.cdlib.org/escholarship>

These institutional repositories contain pre-prints of articles, research reports, conference papers, teaching materials, student projects, theses and doctoral dissertations, committee papers, and

even some copyrighted works that the copyright holder—which may be the publisher—has agreed to have included.

Prepared April 29, 2006