POLICY, TECHNOLOGY, AND THE DIGITAL CORPUS

Digital resources preclude separating technology from library practice. You can't have access without technology, and, in many cases, libraries buy access rather than purchasing resources. Unlike books, digital resources can't be expected to survive through neglect. Without active stewardship, today's digital resources will not survive for decades, much less centuries.

Library policies should mandate effective access to library holdings, collection management that supports a balance of short-term needs and long-term use, and preservation of the cultural record as a basic library mission. When applied to digital resources, those policies require attention to a range of technical issues.

Policy needs to control technology in some areas, such as access versus ownership, supporting open access as a way to improve access and potentially control costs, and establishing and supporting digital-preservation methodologies.

Technology needs to carry out policy in many areas, such as using OpenURL to improve access to licensed material, building the harvesters and indexes that make institutional archives effective ways of improving access beyond licensed materials, and establishing systems such as Lots of Copies Keep Stuff Safe (LOCKSS) (for a brief overview on LOCKSS, see the end of this chapter) to help assure long-term access.

One area of library collections raises a special set of issues: scholarly journals in science, technology, and medicine (STM). Even as the number of STM journals expands faster than most libraries can handle, aggressive price increases based on the individual monopolies of each publication preclude even the wealthiest libraries from maintaining the STM journal collections they might wish to have.

At the same time, most STM journals now offer electronic versions, and an increasing number are likely to shift to pure e-publishing, making the digital version the only version. Without attempting to explore the maze of policy, economy, and technology issues surrounding STM journal publishing, we'll look at one set of potential partial "solutions" to the STM journal crisis: open access and its policy and technology implications.

Collection Policy and Technology

Changes in technology have worked to undermine library collection development policies in three related ways:

- Vastly improved access to full text in digital form, primarily for STM articles, improves immediacy of access and makes resources available around the clock and outside the library's walls. That access comes at a high price, money that typically will come out of the acquisitions budget directly or indirectly.
- Improved funding for research in technology and related fields (that is, STM) leads to more publishing activity, quite apart from a possible

increase in "least publishable unit" publishing (that is, publishing ten slender papers from a research project instead of one major paper). The number of STM journals and the number of articles in each journal continues to expand as a result, at rates considerably greater than increases in library budgets.

- Prices of many (certainly not all) STM journals, particularly for the electronic-only or combined electronic and print versions needed by today's academic libraries, increase at rates substantially higher than the rate of library budget increase.

Barring budget increases at a rate not seen to date (and improbable under current conditions), leaves libraries with three unpalatable options, none of which suits typical policies:

- Eliminating substantial portions of the STM journal literature in order to hold the line on total funding.
- Taking money from monographic, humanities, social science, and other acquisitions budgets, thus reducing the long-term effectiveness of the library collection.
- Increasing the acquisitions budget at the expense of other portions of the library budget, resulting in reduced service or other undesirable outcomes.

No easy or single solution for this ongoing situation exists. Some librarians believed that a shift from print to electronic publishing for STM journals would *itself* ease the budgetary crisis. That no longer seems likely, and in some cases, the shift carries its own long-term negative consequences.

Many initiatives offer hopes of improvement, at least in the long run. These include initiatives to publish lower-cost journals to compete with the most expensive commercial journals (for example, the SPARC initiatives). The highest-profile initiatives today come under the general rubric of Open Access, frequently but not always capitalized.

Open Access and Effective Access

The fundamental principle of open access is that scholarly research articles should be freely available to anyone who can use them, as soon as they are published, at no direct cost to any reader.[1]

Two primary methodologies that support true open access are open access archiving and open access publishing. Both rely on a mix of technologies to support open access as a policy.

Open access publishing, called "gold open access" by some within the open access community, substitutes other means of support in place of subscription charges for online access to refereed articles. Anyone with Internet access may read and print the refereed scholarly articles in any open access journal at no charge and with no permission barriers.[2]

Open access archiving, called "green open access" within the community, involves depositing copies of research papers into digital archives either at the scholar's institution or within multi-institutional, topical, or other larger archives. A deposited research paper may represent the final edited text of a refereed paper published in a traditional journal (identified as such), the paper as originally submitted to the journal, or (when journal policies conflict with open access desires) the paper as submitted, accompanied by a separate list of changes required to replicate the final text.

Open access archives, to be even minimally effective for true access, must adhere to the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) for building metadata on the articles; such archives are called OAI archives or digital repositories. Given that adherence, metadata about articles within an archive can be harvested and combined with metadata from other archives to form large-scale indexes, making the individual archives part of a globally available virtual resource center. Thus, the policy of OAI-PMH adherence and the technology of harvesters combine to carry out the larger policy initiative, making research articles *effectively* available, not just *theoretically* available.

It's important to note the *primary* motivation for most open access proponents is improved access, not relief for library budgets. Some proponents of green open access assume that existing print journals and their publishers will continue along present paths, their profits and prices undisturbed by the added access available via OAI archives.

Given the nonsustainability of that premise, libraries need to find technological means to make both forms of open access more effective, with the goal of eliminating needless, redundant, and overpriced subscriptions.

Open Access and OpenURL

One obvious way to make open access effective access is to build high-quality indexes based on OAI harvesting. The University of Michigan is a leader in this effort with OAIster, its harvesting-based index. As of late 2004, OAIster indexes more than 4.5 million items—a fraction of the STM article literature, but a substantial corpus in its own right.

A scholar within a subject field, however, is unlikely to search OAIster first, particularly if that scholar's library licenses high-quality indexing services in the scholar's own field. The scholar will (and should) start with the subject-specific indexes. How does the scholar get from article citations in those indexes to freely available copies in OAIster?

A similar problem arises for open access journals. They're freely available online, but that means that libraries will not have purchased them. In the past, the act of purchase (or gift exchange or other formal acquisitions' method) triggered cataloging and other steps to make the journal known as part of the library's collection.

Without formal acquisitions, open access journals may not be represented in a library's catalog or journal lists. Index databases already include many high-quality open access journals, but such journals rarely appear in commercial full-text aggregations. How does the scholar get from article citations in subject indexes to the freely available articles in open access journals?

In both cases, one answer may be another technology—one that improves a library's use of its resources by making them more readily accessible. That technology is OpenURL.

OpenURL defines a communications protocol (or choice of protocols) and set of metadata that allow one computer program to communicate selected bibliographic information for a resource to another computer program in an effort to discover availability of that resource.

For example, a researcher using RLG's Anthropology *Plus* database finds an article of interest. If the researcher's library has implemented OpenURL, an icon appears above the article's bibliographic data. Clicking on that icon yields another window that may show that the journal (and volume) in which the article appears is available within one, two, or several different full-text digital resources. Another click on one of those resource links brings up the article itself, assuming the researcher has appropriate permissions.[3]

Good library policy works hand-in-hand with OpenURL technology to make open access effective access.

When libraries assure that high-quality open access journals are reflected in their OpenURL knowledge bases and catalogs, articles in those journals will be as readily available as if the journals were included in commercial full-text aggregations.

When libraries offer links to OAIster or other harvested OAI indexes as additional OpenURL services, researchers may be able to find digital copies of articles in journals to which the library does not subscribe. As OAI indexes grow in size, some libraries may even find it worthwhile to do automatic searches of such indexes as part of OpenURL resolution.

Libraries as Publishers and Related Policy Issues

While open access publishing and open access archiving could relieve library budgetary pressures and improve access to scholarly research, the business model for open access publishing is not as robust as might be desired. It's hard to find a balance among publication (and possibly submission) fees that researchers and their institutions can pay, other reliable sources of income, and the expenses and profit required to sustain a commercial enterprise.

Hundreds of early open access journals, few of which used that name before the twenty-first century, began and continue without publication fees. Many of these journals have been published by universities and, in some cases, university libraries, with the nominal costs of electronic publishing absorbed by the institution as part of its commitment to scholarship.

Libraries may take on larger roles as publishers in the future. Technologically, that's becoming more realistic all the time:

- Mass storage is now "cheaper than dirt" (in Roy Tennant's choice phrase, true enough for some forms of dirt and some forms of mass storage).
- While Internet access is no more free than storage itself, most academic institutions maintain high levels of Internet access for all campus requirements. The bandwidth required to distribute scholarly articles seems likely to be no more than a rounding error in the overall picture of institutional Internet usage.
- E-mail and other digital communications make the refereeing process faster and cheaper. Editing still requires time, but that time may combine voluntary effort and effort supported as part of the scholarly mission.

Overall, academic libraries may find sustainable models for open access publishing more readily than commercial publishers. For that matter, some libraries may already be acting as publishers without the knowledge of the library directors or campus officials.

Such expanded roles raise whole sets of policy issues. The technology is there. Libraries can do this. Does that mean they should? Can a library make the long-term commitment required to establish a prestigious open access journal? Will digital publishing carry out the mission of the library? Feasible doesn't always mean desirable, although such roles may be desirable for some libraries in some fields.

Long-Term Access and Digital Preservation

Nearly all academic and most public libraries include long-term access as a fundamental policy. With non-alkaline printed books, long-term access can sometimes be achieved through neglect: If enough copies exist in libraries around the world, many copies will still be in good shape centuries from now.

It's abundantly clear that digital data isn't forever. No digital medium offers the life expectancy of ink or toner on permanent paper. The usable life spans of most digital media are likely to be much shorter than their physical lives. You might have hard-sectored, eight-inch diskettes containing manuscripts written with The Electric Pencil, and the diskettes may not have deteriorated. Fine: Now how do you read the diskettes and make sense of the files? Digital preservation won't be cheap or easy, even if libraries make the commitment to do it.

One presumption seems to remain constant when digital storage is discussed: Once it's digital, all we need to do is copy it to a current storage mechanism. The information can be copied from medium to medium as often as needed with no possible degradation—after all, "bits is bits." That's a major argument for converting analog resources to digital form, since every time you reformat analog resources, you'll probably lose detail.

Unfortunately, the standing presumption just isn't true, at least not for resources that are "born digital"—things published on digital media or distributed over digital networks. You can't assume that something born digital can be reformatted as often as needed with no loss of content.

On one hand, a growing number of digital publications include digital rights management that precludes easy copying and reformatting. On the other, just copying the data itself—the digital bitstream—doesn't necessarily assure useful access to the resource.

For example, few CD-ROM publications, particularly ones valuable for a library collection, consist entirely of data stored in industry-standard formats. Most good CD-ROM titles combine standard data, proprietary data formats, and software to make it all work properly. The software tends to be system-dependent. Copy it from a Windows platform to a Unix platform, and it doesn't work. In some cases, the software is generation-dependent: It won't work well (or at all) on much newer versions of the same platform.

Additionally, it's possible to produce CD-ROMs and DVD-ROMs that work only if certain *physical* characteristics of the published disc are recognized. Copy the files to any other medium, including a recordable optical disc, and the software just won't work. Such publications can't be reformatted to another medium without ways of undoing this level of protection.

You might assume that none of this matters for "purely digital" resources—those things distributed over the Internet and intended to be used through Web browsers and related programs—that you should always be able to copy such files to current storage systems.

That's not necessarily true. It's trivially easy to encrypt any text or other file, and companies have been offering high-quality encryption schemes for some time now. An encrypted digital object may only become usable when associated software authenticates your right to use it. The software that lets you use the object (text, image, sound, or whatever) almost certainly won't let you save it to disk in clear form. It may not even let you print the text: That's easy enough to prevent, if prevention is desired.

Encrypted files raise two difficulties. The first is that, typically, they place libraries in a pay-per-use situation, in which a fee must be paid each time a resource is viewed. That's a short-term issue, albeit a serious one.

The long-term difficulty is that encrypted files can't be archived locally in any meaningful way. You may be able to save the encrypted file and convert to crystalline storage two generations down the line—but then what? To use the file, you still need authentication from the publisher's or supplier's system.

That becomes tricky if the publisher or supplier goes out of business or the company providing authentication software stops providing it. In that case, authentication won't work—you won't even be able to reach the Internet location—and the file is so much wasted storage space.

Is there a large library that doesn't have books on its shelves from publishers that have gone out of business? Has a library ever been required to destroy books because the publisher's gone under? That's how it works with encryption: no publisher or distributor, no content.

These are all real problems, not straw men. They are not universal problems—and they normally don't affect cases where libraries transform analog resources to digital form. In those cases, the library *should* be able to reformat perfectly. It just takes (lots of) money, people, and ongoing attention to the problem!

Preserving Born-Digital Resources

How do libraries cope with these potholes on the road to digital preservation? While no easy and correct answers exist, some wrong ones have been offered. For example:

- *Rely on the publishers for archiving and reformatting. Assure them they'll sell enough new access to make it profitable.* This is a fundamentally unsound suggestion. While the largest publishers may be around centuries from now and some may find long-term access to be profitable, some of a library's most valuable digital resources will come from smaller, shorter-lived publishers. No publisher usually means no publisher-maintained archive. But then, even surviving publishers won't—and, for stock-owned companies, really can't—maintain long-term access at a loss.
- *Don't buy digital resources that pose such problems.* That's one solution, but it eliminates resources that libraries should have—and librarians may not always *know* whether a resource will pose reformatting problems. It's not reasonable to ignore digital publications and network-based resources, and it's not feasible to require absolute assurances that such resources will be fully archivable.

Work is proceeding. An RLG-OCLC cooperative effort has established guidelines for archival digital repositories.[4] Systems to certify such repositories are on the way. Archival digital repositories won't be cheap, won't be trivial to build, and may not solve the problems wholly. But they do offer the best-known general solution—a combination of policy (the guidelines) and technology (the methodologies)—that may work in most cases.

Sustained Access to Digital Resources

When a library cancels a subscription to a print journal, it retains the previous articles. If the publisher goes out of business, the articles remain intact on library shelves.

When a library cancels its license fee for a full-text aggregation, it generally loses access to all articles in that database, past and future. When a library cancels a subscription to an ejournal, however, it *should* retain access to all the articles it paid for, just as with a print journal. If the ejournal's publisher goes out of business or stops publishing the ejournal, libraries should still be able to use the articles for which they paid. Policy should demand such sustained access, but that policy requires technological backing.

If ejournal contents can be downloaded to a library's or a consortium's mass storage and retrieved from that storage, sustained access should be feasible—but not necessarily assured. Hard disks do crash; backup systems can be corrupted.

One recent initiative may help to assure sustained access as a policy matter, through a combination of innovative technologies. This is the LOCKSS project.

LOCKSS, headquartered at Stanford University, posits a network of institutional mass storage systems (repositories) holding known digital resources with appropriate agreements among the repository owners, publishers with works held on the storage systems, and LOCKSS itself. The copies stored in repositories may or may not be available for patron use while an ejournal (and an institution's subscription) is active; that's also part of the agreements.

LOCKSS establishes a methodology for self-healing repositories. Contents of one repository are checked against the contents of others on an ongoing basis. When discrepancies arise, multiple repositories are examined to assure the discrepancies are corrected. In this manner, lots of copies serve to assure that each copy remains intact—keeping the digital stuff safe.

http://lockss.stanford.edu

Conclusion

Maintaining library policies in an age of ever-growing digital resources requires thoughtful use of technological controls and requires policy oversight over the demands of technology. The general notion that technological solutions can solve all problems caused by technology is nonsense. Within the library field, however, sound policy and innovative technology should ensure that digital resources enhance libraries without undermining traditional roles.

Notes

[1] Some resolutions and advocates add to that definition, requiring that research articles be in such a form that they can be used for data mining. That's an embellishment that won't play into this discussion.

[2] Open access publishing does not require publishing charges paid by authors, their institutions, or research grants, although that's the most

common business model among newer open access journals. Open access journals can and do have subscriptions, either for print versions of articles freely available online or for print journals that add news, editorials, interpretive articles and other non-refereed enhancements to the refereed articles freely available online.

http://cites.boisestate.edu/civ4i2.pdf

[3] That's an extremely brief version of what's actually involved in an OpenURL transaction. A better, if still abbreviated, version appears in *Cites & Insights* 4, no. 2 (Midwinter 2004): 13–14.

www.rlg.org/en/pdfs/repositories.pdf

[4] *Trusted Digital Repositories: Attributes and Responsibilities* (Mountain View, Calif.: RLG, May 2002). Available as a PDF download.