DIGITAL ARCHIVES AND REPOSITORIES *update*

President's Message

Digital Archives and Repositories

For a number of reasons, CRL libraries have come to rely entirely on electronic delivery for many important scholarly source materials. Maintaining costly paper subscriptions to core science and humanities journals produced by Elsevier, Springer, and other publishers is now widely considered an unnecessary expense. And a growing number of libraries no longer subscribe to newspapers in paper or microform, and instead provide users access to electronic news aggregated by vendors like Factiva, ProQuest, and NewsBank.

CRL exists to ensure the long-term availability of critical research materials to member libraries. It has accomplished this by reliably maintaining and delivering newspapers, journals, books, dissertations, archives, and other materials that are not widely held. This strategy does not apply to born-digital research

www.crl.edu

materials. To ensure access to electronic journals, news, and other born-digital content libraries must rely upon whatever archiving arrangements are put in place by producers, publishers, and aggregators. Libraries are often expected to bear the costs of such archiving provisions either directly, by subsidizing the creation and maintenance of digital repositories, or indirectly, through high subscription rates.

Since these costs are significant, due diligence on this front involves determining the extent to which a given repository or archiving arrangement is likely to ensure long-term accessibility of its content. To make such determinations library directors must consider several technical and non-technical factors.

In this issue of *Focus* we identify some of these factors and suggest questions library directors should ask when considering investing in digital preservation solutions. These factors derive in part from the draft *Audit Checklist for the Certification of Trustworthy Digital Repositories* developed by the Research Libraries Group and National Archives and Records Administration. They form the basis of our analysis of a number of digital repositories under the Andrew W. Mellon Foundation-funded CRL Auditing and Certification of Digital Archives project. (We also will test and evaluate specific technical aspects of the repositories.)

As part of the CRL project Robin Dale has profiled three of the digital repositories we are examining. The profile information, disclosed by the repository organizations, will eventually be augmented through on-site technical and operational audits of the repositories. This preliminary information, however, should afford CRL library directors some context for their own risk analysis.

Such analysis and information on digital repositories will enhance the ability of CRL libraries to ensure for their communities long-term access to critical knowledge, as they move with confidence and certainty to efficient forms of knowledge preservation and management.

—Bernard F. Reilly, Jr.

General Factors to Consider in Evaluating Digital Repositories

The increase in the production of and reliance upon electronic resources has intensified the need for "long-lived" digital data.[1] In response to this demand, a variety of repositories for digital content have emerged in recent years. Since these ventures tend to look to the library sector for support, criteria for assessing their reliability are needed. The criteria below were derived from the draft *Audit Checklist for the Certification of Trustworthy Digital Repositories,* developed by the Research Libraries Group and National Archives and Records Administration and from CRL's own work on distributed print archives of journals and government documents.

- Organization—The mission and solidity of the organization that supports the repository will affect the repository's prospects for continuity. Repositories vary, from those created for the express purpose of preserving content for academia to those embedded within scientific, publishing, and aggregator organizations. It is important to know the extent to which preservation is integral to the parent organization's mission, and how important the repository functions are to that organization's revenue stream.

- Governance and Accountability—The governance of the organization that supports the repository determines the communities whose interests will drive the activities of that repository. How accountable is the organization to the user community, and in what ways is that accountability assured? Conversely, how accountable is the organization to the producers or publishing community?

- Content—What content is maintained by the repository and what are its critical characteristics? The extent and scope of the journal titles, databases, and other materials archived should be listed, or easily discovered, and verifiable. What mechanisms are in place to ensure the continued deposit of the listed content and prevent its withdrawal by the publisher?

- Ingestion—Trustworthy repositories will disclose specific data on the form and functionality of the content ingested. Most archives reformat or "normalize" content in order to limit the cost of managing and migrating complex formats. Normalization may make the archived content look or behave differently than it does when delivered directly to users by producers or publishers. Clarity about the nature and degree of normalization can provide a sense of the scale of investment the library and/or the repository will have to make, if any, to provide an acceptable level of functionality in the future.

[1] The call for persistent collections has come from sectors as disparate as the Andrew W. Mellon Foundation, in Donald J. Waters' October 2005 paper *Urgent Action Needed to Preserve Scholarly Electronic Journals,* and the National Science Board with the May 2005 report *Long-Lived Digital Data Collections: Enabling Research and Education in the 21st Century.*

- Technical Systems and Data Security—The most obvious indicator of the reliability of a repository is the stability and robustness of its technical infrastructure. Factors here include whether or not the repository system conforms to the Open Archives Information Systems Reference Model, to various system security requirements and standards developed in government and other domains, and whether or not the policies and methods for backup, redundancy, authentication, and distribution of functions and services are clear and conform to accepted best practices. Also important is the scalability of the system. Is the repository likely to be able to accommodate new and complex forms of content and functionality?

- Cost Structure and Distribution—The costs of a repository can be structured and distributed in several ways, with differing implications for future costs to the library. The repository may assess the library or users a combination of initial capital fees and ongoing maintenance fees, or simply a subscription fee. Some costs might also be borne by the publisher of the archived content. While there are limits to how precisely a repository can project future fees in advance, libraries should be clear about the cost drivers (such as amount and complexity of content, frequency of migration, royalties to content publishers, etc.) and how the costs are distributed in the event of changes in those drivers.

- Rights—Repositories should disclose documentation of the rights they hold to deliver the content in the event of failure by the producer or publisher, the duration of the grant of those rights, and whether said rights are transferable. Such documentation should be clear about what constitutes failure. Failure is often defined as when a publisher no longer offers the content, but drastic subscription price increases, the decision to make the content available only as part of a larger, prohibitively priced bundle, and similar events can also put content out of reach of libraries.

- Results and Outputs—Longevity and performance are important indicators of the reliability of a repository. While digital preservation is only just emerging, organizations and systems that have proven histories of effectively fulfilling preservation functions are likely to continue to support persistence.

We will refine and expand these criteria as the Auditing and Certification of Digital Archives project progresses. For more information visit the CRL project Web site. ✥

Profiles of Some Existing Digital Repositories

The common characteristic of digital repositories is that they exist to preserve electronic data and content. But existing repositories of scholarly source materials vary widely in the kinds of content they preserve, the activities they perform, and the funding, governance, and organizational models they adopt. Robin Dale, project director of the Auditing and Certification of Digital Archives project, has developed detailed profiles of two major but quite different digital repositories: the repository of datasets and collections maintained by the Inter-university Consortium for Political and Social Research (ICPSR) and the e-Depot, an electronic journals archive developed and maintained by Koninklijke Bibliotheek in the Netherlands. These profiles can be downloaded from the project Web site at:

- http://www.crl.edu/content/DigArc/DigArc2/ICPSR.pdf (ICPSR) and

- http://www.crl.edu/content/DigArc/DigArc2/KBprofile.pdf (E-Depot).

The profiles provide information about the organizational structures, systems architecture, funding systems, and user communities of those repositories, as well as about agreements between repositories and the publishers or producers that deposit data and electronic content. The profile information is based on pre-audit analysis of the repositories and on documents and information disclosed by the parent organizations. It provides a starting point for further CRL analysis and auditing to determine the relative reliability of the various repository solutions that are now emerging.

An additional profile for the emerging repository Portico, is also available at http://www.crl.edu/content/DigArc/DigArc2/portico.pdf. ◈

The **Center**
for **Research**
Libraries

MISSION STATEMENT

The Center supports advanced research and teaching in the humanities, sciences, and social sciences by ensuring the survival and availability of the knowledge resources vital to those activities.