Learned Publishing, 21, 285–294
doi:10.1087/095315108X356716

Digital preservation: challenges and implementation

Amy J. KIRCHHOFF
Portico

ABSTRACT. The research of the future requires access to the research of the past. This access cannot be assured without reliable long-term preservation of scholarly digital content. Near-term access can be guaranteed with backup and access system redundancy. Mid-term access can be protected with byte replication. But assurance of long-term access requires digital preservation – the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability, and accessibility of content over the very long term. Portico, with a mission to preserve scholarly digital content, is one organization providing such long-term digital preservation.

Introduction

The amount of digital-only content, and the number of its users, have both increased rapidly over the past 25 years from the advent of the first IBM PC in 1981, through the release of the World Wide Web in 1992, and to the Web 2.0 mash-up we know today in 2008. The Internet has made it possible for anyone to be an author. According to WorldWideWebSize.com, the 'indexed Web' contains at least 27.85 billion pages as of June 2008 on an estimated 168,408,112 sites; 2.7 million sites were added in May 2008 alone.[1]

The amount of scholarly digital content has also grown, and this increase is reflected in the growing percentage of academic library expenditures devoted to electronic resources. Between 1993 and 2006, electronic materials expenditures at the libraries of the Association of Research Libraries (ARL) have increased over five times faster than total library materials expenditures.[2] In 2005–6, these libraries spent an average of 41% of total library materials expenditures on electronic resources, and 23 ARL libraries spent more than 50% of their materials budget on electronic resources (Figure 1).

Over the past decade, published scholarly literature in digital form has grown, reliance on this content by the academic community has grown, and expenditures for this content have grown. Yet digital content is ephemeral. The first full-text search engine, Web Crawler, debuted in 1994 and indexed approximately 72,000 pages. None of the top 25 pages listed at that time exists today. Without measures to ensure the long-term preservation of e-journals, we have no assurance that, a generation from now, today's e-journals will not suffer a similar fate. Digital preservation is needed to ensure that future scholars will be able to access and build upon today's research and science.

Amy J. Kirchhoff

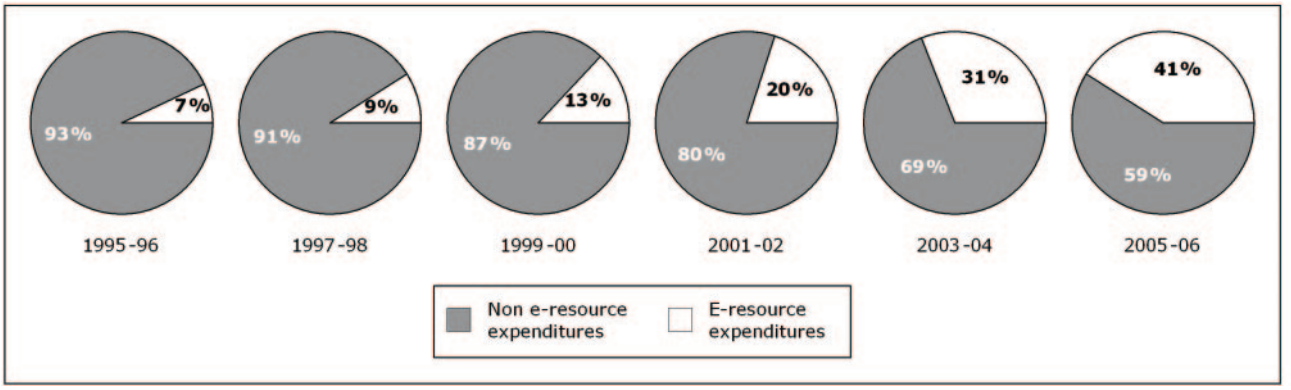

Figure 1. Library materials expenditures 1995–2006 (from *ARL Statistics*)[3]

the potential loss of e-journals is unacceptable

Technology and file formats will evolve, and today's knowledge will be lost over time without special preservation intervention.

In September 2005, the need for a robust archiving solution for e-journals was expressed in the *Urgent Action Needed to Preserve Scholarly Electronic Journals* statement which was endorsed by the Association of Research Libraries, the Association of College and Research Libraries, and others.[4] As a follow-up to this call to action, Portico and Ithaka undertook a survey in early 2008 of library directors of four-year colleges and universities in the United States to examine current perspectives on the preservation of e-journals. The survey found that a large majority of library directors across the spectrum strongly agreed or agreed that the potential loss of e-journals is unacceptable, and a significant majority believed their own institutions have a responsibility to take action to prevent an intolerable loss of the scholarly record.[5] Amidst this acknowledged need, reliable long-term preservation arrangements for e-journals and other scholarly literature published in electronic form have been emerging.

Near-term access protection and long-term content preservation

The methods a gardener uses to 'preserve' strawberries for the coming winter months are quite different from those a plant biologist uses to save specimens for study over the coming decades. So, too, the methods the

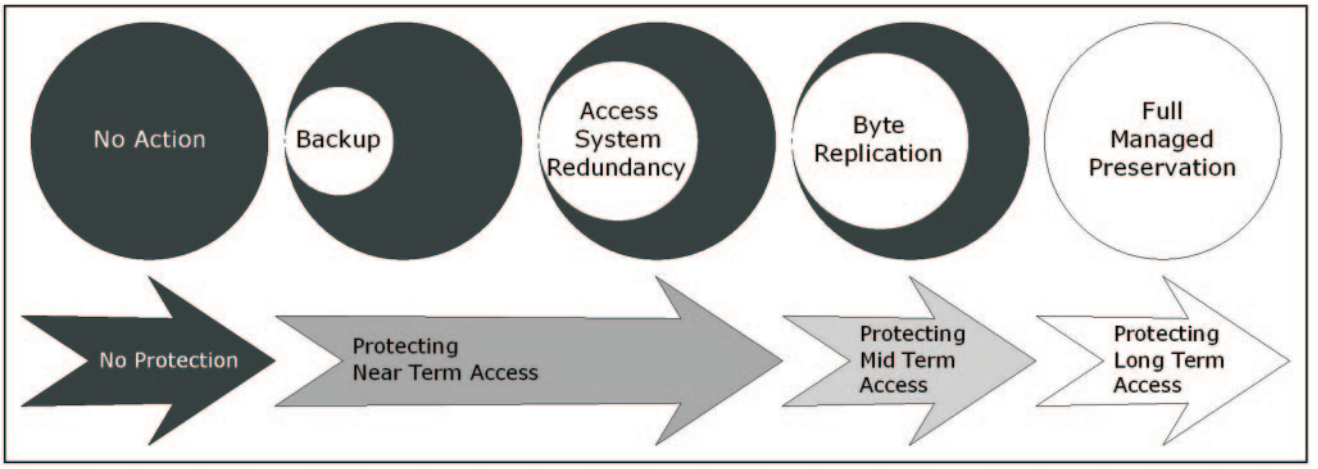

Figure 2. Near-term access protection and long-term content preservation

scholarly community uses to protect content for use in the near term differ from those used to preserve content over the long term. The steps involved in both near-term access protection and long-term content protection through preservation can be placed along a continuum (Figure 2).

Digital preservation is the series of management policies and activities necessary to ensure the enduring usability, authenticity, discoverability and accessibility of content over the very long term.[6] The key goals of digital preservation include:

- *usability* – the intellectual content of the item must remain usable via the delivery mechanism of current technology;
- *authenticity* – the provenance of the content must be proven and the content an authentic replica of the original;
- *discoverability* – the content must have logical bibliographic metadata so that the content can be found by end-users through time; and
- *accessibility* – the content must be available for use to the appropriate community.

The early points on the continuum that protect near- and mid-term access to content – backup, access system redundancy, and byte replication – do not provide long-term digital preservation.

Backup

Backup, when content is copied and stored in multiple locations to create readily available data replacements in case of equipment failure or other catastrophe, has long been understood to be a requirement for protection of near-term data access. It is imperative for business continuity, and it is necessary to ensure that access to content in the near term will not be interrupted for any length of time. A well-managed backup system can help quickly resolve problems with content encountered this week, or next week, or next month, but not over the long term. Backup is typically implemented with commercial software that allows users to retrieve files backed up at specific points in time. Very often, content may only be retrieved via the software with which it was originally backed up. If special software or hardware is required to access the content and if it has been compressed via a proprietary technology, the long-term future accessibility and authenticity of the content – key goals of digital preservation – cannot be assured.

Access system redundancy

Many content delivery systems are configured for redundancy, so that the entire system is running over two or more computers in two or more data centers. These multiple systems may be online at the same time, or one may be a 'hot spare' that can quickly be brought online should the first system fail. Access system redundancy is an excellent way to ensure that there is little interruption to near-term, ongoing access, but it does not alone guarantee usability, authenticity or accessibility of the content over the long term as technology and data formats evolve.

many content delivery systems are configured for redundancy

Byte replication

Byte replication is a process whereby identical, multiple copies of files, file systems, or websites are created. They may be written to other online computers or to offline media. These replicas are typically held in diverse geographic locations and specialized software is not needed to access the content. This diversity in copies and location, together with the lack of reliance on software, ensures that byte replicas should provide content that is authentic and usable for as long as the file formats remain readable. However, simple byte replication includes no provision for ensuring the content is usable when the file formats are no longer current, nor is there any inherent provision for ensuring that the content remains discoverable. For example, if a series of e-journal files are byte-replicated, without accessible bibliographic information describing the intellectual content of the replica, there is no guarantee that an end-user in the future will be able to find the specific article he or she needs.

Digital preservation

Digital preservation is the series of manage-

ment policies and activities necessary to ensure the enduring usability, authenticity, discoverability, and accessibility of content over the very long term. While backup, system redundancy, and byte replication may be used by delivery organizations and digital preservation organizations, these actions alone are not sufficient for digital preservation. The following components are necessary to achieve digital preservation:

backup, system redundancy, and byte replication are not sufficient for digital preservation

An independent organization with a mission to carry out preservation

As noted in a recent CLIR survey, *E-Journal Archiving Metes and Bounds: A Survey of the Landscape,* the first indicator of an archiving program's reliability is that it 'have both an explicit mission and the necessary mandate to perform long term e-journal archiving'.[7] The mission creates an environment conducive to the specialized planning and infrastructure needed to support digital preservation. In addition, placing the content in a third-party environment, separate from both the original content creator and the content consumer, requires that the preservation organization demonstrate a capacity to support the content in systems separate from those that originally created and sustained it.

A sustainable economic model that can support preservation activities over the targeted timeframe

The actual costs of long-term digital preservation are difficult to determine with accuracy at this early date; however, the JISC LIFE Project[8] proposes that there will be an ongoing 'technology watch' cost that is relatively stable year to year, and an intermittent set of costs related to the taking of 'preservation actions' (Figure 3).

A preservation organization may choose to preserve all content in its archive for the same amount of time, or may preserve different items for different lengths of time. The retention time of every item preserved should be set at the point the content is acquired and imported into the archive. The preservation organization must have an economic model that provides enough funds to move the content into the archive, continue its ongoing technology watch, and implement intermittent preservation activities on the preserved content for the stated retention time period.

Clear legal rights to preserve the content and relationships with the content providers

A digital preservation organization must have legal rights to preserve the digital content ingested into its archive. These rights must be obtained in advance of the acquisition of the content and must delineate the scenarios under which the preserved content is to be made accessible and to whom; the rights must include the ability of the preservation organization to delegate a successor, should the original organization cease operations. Because these legal agreements must always remain with the archived content,

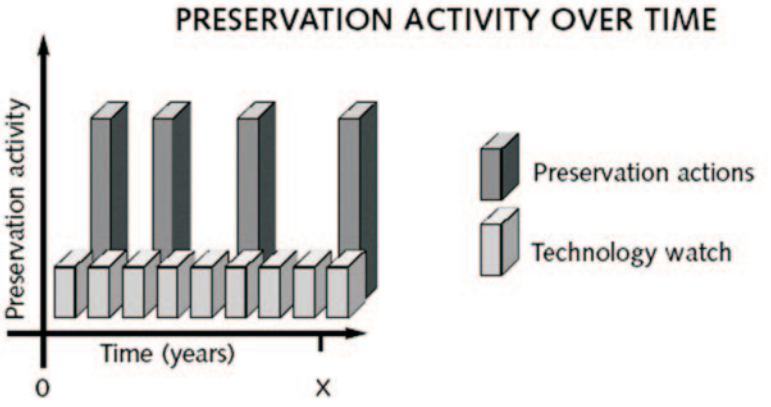

Figure 3. LIFE Model – preservation activity over time.[13]

they should be preserved in the archive along with the digital content.

In addition to obtaining the legal rights to preserve the content and to make it available under specified conditions, the preservation organization must have a relationship with the content provider. Two of the key elements of digital preservation are that the content must remain usable and accessible over the long term. In order to fulfill these requirements, the preservation organization must understand the structure and format of the content. For example, while staff at the preservation organization need not understand physics to preserve physics journals, the staff must understand how scholarly journals – and physics journals in particular – are typically structured, and how a particular journal is structured. With this knowledge, the preservation organization can develop the tools necessary to automatically read the source files that comprise the journal and to retrieve the bibliographic data – the data which ensures that the content is discoverable in the future. Reaching this level of understanding of the content structure often requires interactions with the content provider early in the preservation lifecycle.

Relationships with the eventual users of the content

As noted in the OAIS standard (discussed below), an archive exists to support a designated community, 'an identified group of potential consumers who should be able to understand a particular set of information. The designated community may be composed of multiple user communities.'[9] The preservation organization must have an ongoing relationship with its designated community and must have a way of eventually delivering usable, authentic, and discoverable content to that community.

A preservation strategy consistent with best practices and a technological infrastructure able to support the selected preservation strategy

Best practices and standards in digital preservation are continuing to evolve, but a number of guidelines do exist, including:

- OAIS (Reference Model for an Open Archival Information System, ISO 14721: 2003) – a high-level framework for designing a preservation organization;[14]
- PREMIS (Preservation Metadata: Implementation Strategies) – a data dictionary and documentation describing the metadata necessary for preserved digital content;[10]
- TRAC (Trustworthy Repositories Audit & Certification: Criteria and Checklist) – a set of digital preservation best practice criteria that can be used to evaluate repositories;[11]
- DRAMBORA (Digital Repository Audit Method Based on Risk Assessment) – a risk management methodology that allows a repository to run an internal audit in order to assess its capabilities, weaknesses, and strengths;[12]
- nestor (Network of Expertise in Long-Term Storage of Digital Resources) Catalogue of Criteria for Trusted Digital Repositories—a checklist similar to TRAC to assess the technical and organizational trustworthiness of a digital repository;[13]
- DPC Handbook (Digital Preservation Coalition) – a detailed guide to the management of and long-term access to digital objects.[14]

the content must remain usable and accessible over the long term

Transparency about preservation services and strategies, clients, and content

Just as the OAIS standard details how an organization should define its designated community and its clients, the CLIR report[12] details a set of indicators of reliable digital preservation repositories. The report notes that a preservation organization must 'be explicit about which scholarly publications it is archiving and for whom [and] offer a minimal set of well-defined archiving services'.[12] A preservation organization should clearly communicate its preservation methodology to its designated community and provide the community with a method of auditing the preserved content.

Technical methods of digital preservation

Migration and emulation are the two primary strategies used for long-term preser-

vation. Migration involves transforming digital content from its existing format to a different format that is usable and accessible on the technology in current use. Emulation involves developing software that imitates earlier hardware and software. Migration is a strategy that requires a deep understanding of the content being preserved, whereas emulation is a more technology-based strategy, requiring a deep understanding of existing hardware and software.

Emerging organizational models for digital preservation

the different technical strategies for preservation can be implemented within different organizational models

The different technical strategies for preservation can be implemented within different organizational models. Three models are broadly acknowledged today for digital preservation of scholarly content:

1. *National libraries.* A number of national libraries have taken on digital preservation in support of their mission or their country's legal deposit requirements. The scope of content involved and access terms vary, but all such libraries are government-funded. Examples of this type of organization are the British Library[15] and the National Library of the Netherlands.[16]
2. *Community-supported independent preservation archives.* These organizations may focus on a subject area or content type. Typically the costs for the preservation of this content are shared across the participating publishers and libraries. Examples of this type of organization are Portico[17] and the Inter-University Consortium for Political and Social Research (ICPSR).[18]
3. *Networked library efforts.* Groups of libraries have pooled their resources to share the responsibility and costs of preservation. Examples include LOCKSS (Lots of Copies Keeps Stuff Safe),[19] CLOCKSS (Controlled LOCKSS),[20] and NDIIPP (National Digital Information Infrastructure Preservation Program – the digital preservation program of the US Library of Congress).[21]

Case study: Portico

Portico is a not-for-profit organization with a mission to preserve scholarly literature published in electronic form and to ensure that these materials remain available to future generations of scholars, researchers, and students. With support from JSTOR, Ithaka, The Andrew W. Mellon Foundation, and the Library of Congress, Portico was officially launched in 2005 with an initial focus on e-journal preservation. In response to the growth of digital content of all genres and types, and the community's expressed desire to have a greater portion of this important content preserved, Portico is now expanding its preservation work to include e-books, digitized newspapers, and libraries' locally created or digitized content.

Libraries and publishers are the two key groups of participants in the Portico preservation service. Libraries have traditionally held the responsibility for archiving the scholarly record. Libraries would purchase print content to make it accessible to their communities and they would then preserve that content in the library stacks. In the digital world, however, libraries do not actually 'receive' a copy of the digital content; neither do they, individually, have the infrastructure necessary to support receiving and hosting a copy of all the digital content which they license. Publishers have additional preservation responsibilities with digital content, compared with print, because the digital files remain in their hands and are therefore their responsibility. Publishers and libraries both participate in Portico to meet some of their preservation responsibilities through their support of third-party preservation of digital scholarly content. Publishers are responsible for providing their digital files to Portico, and for making an annual financial contribution to the preservation service. Libraries are also responsible for making an annual financial contribution and for auditing the archive. Library participants at Portico may designate up to four auditors who have access to an audit interface into the archive. Through the audit interface, the library auditors can check that content is being added to the archive and preserved. Libraries and publishers participating in the Portico e-journal preservation offering both agree that e-journals in the archive will become broadly accessible to the faculty, staff, and students

at the institutions of participating libraries via a delivery site, under certain specific scenarios. This broad access occurs in the case of a trigger event (when a publisher ceases operations, ceases to publish a title, no longer offers back issues, or suffers catastrophic and sustained failure of its delivery platform) or in the case of a post-cancellation access request (some publishers choose to designate Portico as one method to provide post-cancellation access to their library subscribers' content).

The Portico preservation archive is compliant with OAIS, the Reference Model for an Open Archival Information System. The Portico archive is technology- and application-independent (e.g. the Portico archive can be exported into a standard file system with all the information necessary to understand the contents of the archive in organized files). Portico's digital preservation service uses a migration-based preservation strategy. Portico will 'migrate' or transform the preserved content from one file format to another as technology changes. Portico supplements and supports this migration policy by preserving the original source files along with all migrated versions.

Portico's digital preservation service includes:

1. *Preservation planning*. Portico analyzes the formats and packaging of the supplied content and develops a preservation plan appropriate to the content and to the needs of the publisher and library participants in the service, and guided by Portico's established preservation policies. This specific preservation plan may include an initial migration of the content to archival standards.
2. *Receipt and inventory management*. Portico receives content in a variety of ways including on portable media (e.g. tape or disk), via a standard transfer protocol (e.g. FTP or OAI-PMH/ORE[22]), or via Portico-developed software integrated into digital collections' management platforms such as DSpace[23] or Fedora.[24]
3. *Processing and archival deposit*. Portico will ingest content into the archive according to the specific preservation plan. To enable detection of and recovery from loss or damage to the preserved content, Portico replicates the archive onto diverse media and in multiple geographic locations.
4. *Monitoring and management*. Portico will perform regular fixity and completeness checks of preserved content and restore any content identified as damaged from an archive replica. A fixity check compares the exact bytes being used by a file today with the bytes used by that file in the past and reveals any corruption. A completeness check compares the contents of the entire archive today with the contents of the archive in the past and reveals any content loss. As technology changes, Portico will modify preservation plans and migrate preserved content to appropriate new media and formats to ensure ongoing usability and authenticity. Portico seeks regular accreditation from community-approved archive audit agencies, such as the Center for Research Libraries (CRL),[25] and makes preserved content available for inspection by library and publisher participants. With the e-journal preservation service, for example, libraries may audit all preserved e-journal content and publishers may audit the content they have provided.

e-journals in the archive will become broadly accessible under certain specific scenarios

To facilitate transparency, Portico will provide to its participating publishers and libraries (as appropriate):

- *Documentation*
  - Preservation policies;
  - Reports on the status of one-time and ongoing receipt and processing of content;
  - Annual status reports on preserved content, including detailed holdings lists, repair reports, and migration reports.
- *Content*
  - Audit and inspection access to the preserved content;
  - A copy of the preserved content in its archival format, upon request of the owner.

The structure of Portico's preservation planning and specific preservation actions is shown in Figure 4.

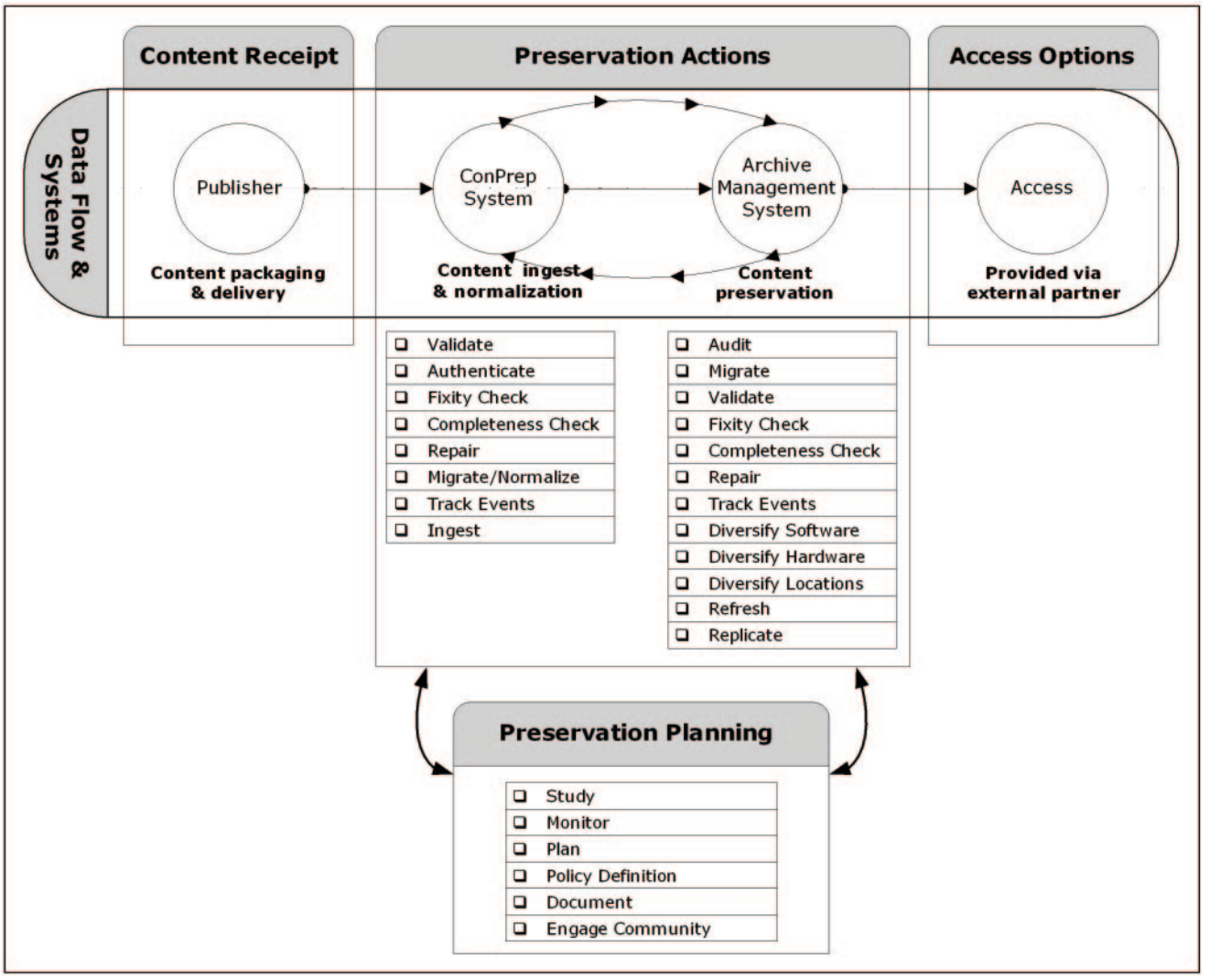

Figure 4. Portico's managed preservation process.

The first step in preservation of content at Portico is planning. When content is initially received, we work closely with the content provider (in the case of e-journals, this is typically the publisher or the publisher's vendor) to gain a deep understanding of the specific content to be preserved. With this information in hand, Portico develops specific policies and procedures for the content, informed by our overarching preservation philosophy and the needs of the content and the content owner. Preservation planning is an ongoing process, and Portico continually engages with the community to understand the current state of best practices in digital preservation, as well as the state of individual file formats and their current life expectancy.

Once an initial preservation plan for the specific content has been developed, Portico then turns to the set of preservation actions necessary to implement the plan. Numerous activities are required at the beginning of the preservation process to facilitate the ingestion of the content into the Portico archive. One typical activity is the migration of the content packaging to the Portico data model. Content comes to Portico structured in many different ways – one publisher may compress all the content together into one

file and another may send full directory systems; one publisher may put all the content for an issue into one directory and another may break an issue down into multiple directories. Portico analyzes this content packaging and writes a tool for each publisher, to transfer each unique packaging system into the Portico data model. Another typical initial activity is the migration of files to archival formats. For example, publisher XML files may be migrated to the NLM Journal Archiving and Interchange DTD[26] and individual TIFF page images may be migrated to a single PDF file. Portico preserves the original files and the migrated files in the archive. After content has been deposited into the Portico archive, the archive is replicated, and the replicas are regularly monitored and compared to check for any inconsistencies, which would imply data loss or corruption. Content is repaired as needed over time. When necessary, copies of the content will be extracted from the archive and reprocessed for migration to new formats. The migrated content will then be deposited into the archive, together with the original archival packages.

Portico has managed an operational archive since March 2006. As of June 2008, Portico has 56 participating publishers that have committed over 7,600 journals to the archive, and 465 participating libraries from 13 countries. Portico's financial model assumes that over time the primary beneficiaries of the e-journal preservation service – publishers and libraries – will provide the primary financial support for the archive, with additional support provided by charitable foundations or government agencies. In 2007, 55% of support was provided by charitable foundations, with libraries contributing 35% and publishers 10%. As of June 2008, Portico had preserved over 7.35 million articles in the archive or over 77.75 million individual files, and the archive is currently approximately 6.6TB in size. The Portico preservation service has systems capacity to process approximately 60,000 articles per day (or 1–2 million articles per month).

Conclusion

The world is going digital at an astounding rate. Desktop computers were available only in 1981; today, a mere 27 years later, desktop computers and the Internet are ubiquitous. The scholarly community, like the general public, is moving rapidly into this digital environment, and the digital scholarly content being created today must be preserved for the future – without guaranteed access to this record over the very long term, future research will be hampered. Traditional preservation responsibilities and methodologies are not applicable to digital content, where physical copies are not delivered to libraries. Instead, in order to meet the unique preservation needs of digital content, libraries and publishers – two key participants in the scholarly communications environment – must join together to 'invest in a qualified archiving solution'.[4] Portico, as a third-party, community-based preservation service, is one such qualified archiving solution.

traditional preservation responsibilities and methodologies are not applicable to digital content

References

1. Netcraft. May 2008 Web Server Survey, 2008. http://news.netcraft.com/archives/2008/05/06/may_2008_web_server_survey.html (Accessed Jun 3, 2008).
2. Kyrillidou, M. and Young, M. *ARL Statistics 2005–06: A Compilation of Statistics from the One Hundred and Twenty-Three Members of the Association of Research Libraries*. Washington, DC, Association of Research Libraries, 2008. http://www.arl.org/stats/annual surveys/arlstats/arlstats06.shtml (accessed 22 May 2008).
3. See http://www.arl.org/stats/annualsurveys/arlstats/index.shtml for ARL statistics over time.
4. Waters, D. (ed.). *Urgent Action Needed to Preserve Scholarly Electronic Journals*. Digital Library Federation, 2005. http://www.diglib.org/pubs/waters051015.htm (Accessed 1 May 2006).
5. *Digital preservation of e-journals in 2008: Urgent Action revisited – Results from a Portico/Ithaka Survey of U.S. Library Directors*. New York, NY, Portico & Ithaka, 2008. http://www.portico.org/comment/ (Accessed Jun 22, 2008).
6. This definition is based on definitions found in *Trusted Digital Repositories: Attributes and Responsibilities*. Mountain View, CA, Research Libraries Group-OCLC, 2002 (http://www.rlg.org/longterm/repositories.pdf) and Beagrie, N. and Jones, M. *The Preservation Management of Digital Material Handbook*. York, Digital Preservation Coalition with the National Library of Australia and the PADI Gateway, 2002, http://www.dpconline.org/graphics/handbook/
7. Kenney, A.R., Entlich, R., Hirtle, P.B., McGovern, N.Y., and Buckley, E.L. *E-Journal Archiving Metes and Bounds: A Survey of the Landscape*. Washington, DC, Council on Library and Information Resources, 2006. http://www.clir.org/pubs/abstract/pub138abst.html (accessed May 29, 2008).
8. McLeod, R., Wheatley, P., and Ayris, P. *Lifecycle information for e-literature: a summary from the LIFE project*. Presented at the LIFE Conference, April 2006.

http://eprints.ucl.ac.uk/1855/ (accessed 29 May 2008).
9. *Reference Model for an Open Archival Information System – OAIS*. Washington, DC, National Aeronautics and Space Administration – Consultative Committee for Space Data Systems, 2002. http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed 2 May 2006).
10. PREMIS (Preservation Metadata: Implementation Strategies) Working Group. OCLC, 2005. http://www.oclc.org/research/projects/pmwg/ (accessed 19 June 2006).
11. *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist*. Chicago, IL, Center for Research Libraries and OCLC, 2007. http://www.crl.edu/content.asp?l1=13&l2=58&l3=162&l4=91 (accessed 22 June 2008).
12. DRAMBORA, http://www.repositoryaudit.eu/ (accessed 22 June 2008).
13. *nestor Catalogue of Criteria for Trusted Digital Repositories*. Frankfurt am Main, Germany, 2006. http://www.dcc.ac.uk/tools/nestor/ (accessed 22 June 2008).
14. Beagrie, N. and Jones, M. *The Preservation Management of Digital Material Handbook*. Digital Preservation Coalition with the National Library of Australia and the PADI Gateway, 2002. http://www.dpconline.org/graphics/handbook/ (Accessed 28 May 2008).
15. British Library – Digital Preservation, http://www.bl.uk/dp (accessed 22 June 2008).
16. Koninklijke Bibliotheek – National library of the Netherlands: e-Depot and digital preservation, http://www.kb.nl/dnp/e-depot/e-depot-en.html (accessed 22 June 2008).
17. Portico, http://www.portico.org (accessed 22 June 2008).
18. ICPSR: Inter-University Consortium for Political and Social Research, http://www.icpsr.umich.edu/ (accessed 22 June 2008).
19. LOCKSS, http://www.lockss.org (accessed 22 June 2008).
20. CLOCKSS, http://www.clockss.org (accessed 22 June 2008).
21. The Library of Congress – Digital Preservation, http://www.digitalpreservation.gov/ (accessed 22 June 2008).
22. OAI-PMH and OAI-ORE are maintained by the Open Archives Initiative, http://www.openarchives.org/ (accessed 22 June 2008).
23. DSpace, http://www.dspace.org/ (accessed 22 June 2008).
24. FedoraCommons, http://www.fedora-commons.org/ (accessed 22 June 2008).
25. Center for Research Libraries, http://www.crl.edu (accessed 22 June 2008).
26. See http://dtd.nlm.nih.gov/

Amy J. Kirchhoff
Archive Service Product Manager, Portico
100 Campus Drive, Suite 100
Princeton, NJ 08540, USA
Email: amy.kirchhoff@portico.org