# Preserving Content: A Case Study of a Multi-Faceted Approach

*Eileen Gifford Fenton*
*Electronic-Archiving Initiative, JSTOR*
*Princeton, New Jersey, USA*

## Abstract

Fundamentally, preservation is about ensuring the ongoing usability of valued content. The media which convey this content to a reader--paper, microform, CD-ROM, or online product--continue to change over time. But regardless of media variation, the fundamental goal of preservation does not change; the goal remains ensuring the longevity of the content. The challenge is to ensure that there are sufficient provisions and support for the successful transition of the content from one medium to another in a way that is compatible with the goal of long-term preservation.

The likelihood of meeting this challenge is increased through a multi-faceted approach implemented within an organizational framework designed to provide the necessary support for the work of preservation. JSTOR, a not-for-profit organization with a mission focused on the long-term preservation of and access to scholarly resources, provides an opportunity to examine how content preservation and media transitions can be conducted within an organizational framework designed to robustly support this work. Drawing upon her experience as the former director of JSTOR's production unit and the leader of its Electronic-Archiving Initiative, the author suggests several elements useful for preserving journal literature in various media and highlights how these elements may be combined with key organizational components in order to meet the challenges of long-term content preservation.[1] The preservation components discussed include stewardship of print originals, migration of print content to digital media, and preservation of electronic publications. These components by themselves, however, are insufficient to ensure the preservation of content; they will lead to preservation only when developed within an organizational context which includes an institutional commitment to the preservation goal; a business model which can ensure the sustainability of the archival operation; a technical infrastructure able to meet the demands of the content; and appropriate relationships with libraries and content owners and producers.

## Introduction

The medium by which important scholarly content is conveyed to readers continues to change over time. But the desire to preserve this content and to ensure that it will be accessible to future generations of students, researchers and scholars prevails regardless of the medium which originally carried the content. Preservation must succeed whether the content was originally presented via paper, microform, CD-ROM, or online product. The challenge is to ensure that there are sufficient provisions and support for the successful transition of the content from one medium to another in a way that will meet the needs of generations to come.

There is no single "correct" way to meet this challenge, but the likelihood of success is increased through a multi-faceted approach implemented within an organizational framework designed to provide the necessary environment and support for the work of preservation or archiving. "Archiving," for purposes of this discussion, is understood to mean the work necessary to ensure the longevity of and ongoing access to important scholarly resources over a very long time horizon (75+ years). JSTOR, a not-for-profit organization with a mission focused on the long-term preservation of and access to scholarly resources, provides an opportunity to examine how content preservation and media transitions can be conducted within an organizational framework designed to robustly support this work.

JSTOR has created and is maintaining a trusted digital archive of the full back runs of academic journals. Through the development and enhancement of this searchable, interdisciplinary collection, JSTOR's objective is to help all participants in the scholarly community--libraries and publishers, faculty and students--to be more productive, while simultaneously reducing system-wide costs and increasing convenience.

The diversity of journal literature entrusted to JSTOR's archival care highlights the need for a multi-dimensional approach to preservation. Because journals are now routinely presented via both print and electronic media, robust preservation of the content may well require several inter-related components. These might include: a) stewardship of print originals; b) migration of print content to digital media; and c) preservation of electronic publications. These components, which are described below, must be implemented within an organizational context that incorporates at least five elements that are critical for any long-term archive. These include: 1) an institutional commitment to the preservation goal; 2) a business model which can ensure the sustainability of the archival operation; 3) a technological infrastructure able to meet the demands of

the content over time; 4) appropriate relationships with libraries that in turn represent the interests of the scholars; and 5) appropriate relationships with content owners and producers.

## Preservation Components

**Stewardship of Print Originals: Paper Repositories**

College and university libraries have for decades operated as an informal network through which many, many copies of scholarly journals were housed and preserved. No single institution retained a full run of all journals, but taken as a whole the system was populated with enough individually collected and maintained copies that sufficient redundancy seemed to be achieved. Each institution, through cooperative arrangements such as interlibrary loan and document delivery services, could be assured of access to a comprehensive collection of scholarly journals, and so the overall needs of the community were adequately met.

Over time, however, individual library collections do change in response to a number of factors such as shifting local curriculum needs, space constraints, or budget limitations. Consequently journal subscriptions are terminated as budgets are constrained or journals may be deaccessioned as disciplinary focus shifts. Quite apart from deliberate changes to a library's collections, usage imposes unplanned changes over time. Journal issues are lost or damaged or simply wear out as a result of extended use. Collections may also change as libraries choose to free up precious on-campus shelf space by moving to off-site storage those print materials that have been digitized and archived by a trusted third party. In some cases these materials may be deaccessioned. The collective effect of these factors is to gradually reduce the number of print copies of important scholarly journals retained throughout this informal network. While this has the positive effect of reducing the total dollars that are spent system-wide on maintaining a vast number of duplicate copies, it does raise the possibility that over the long term fewer and fewer copies will be safely maintained.

While this trend toward fewer copies may still be in the very early stages, there is a need to ensure that multiple complete paper copies will remain safely preserved for the long term. JSTOR is working with cooperating libraries and library organizations to ensure that multiple copies are safely housed in geographically distributed locations. These copies, which are not available for regular use, form a safe network of copies of last resort. Similar local and regional repository efforts are also beginning to take root and will add an additional level of security for the print originals. Taken as a whole, this formal network of reliably preserved paper copies complements the long standing informal network, and provides one key preservation component.[2]

**Migration through Digitization**

Migration is widely recognized as a crucial part of any successful long-term preservation plan,[3] but typically migration refers to those measures necessary to ensure that an archive and the content under its care keeps pace with changing technologies. It is useful, however, to think of migration more broadly. As the scholarly community has come to increasingly demand the access and convenience offered by electronic resources, it has become almost mandatory to "migrate" print materials to an electronic format. In some circles there is a sense that if it is not digital, it does not exist. The challenges of media migration are very similar to those of technology migration, and this overlap provides a useful way to consider the second important preservation component: migration through digitization.

Migration through digitization may be the preservation component for which JSTOR is best known. JSTOR migrates content from its original paper media to digital media by creating a digital image of each printed page, a separate image for grayscale or color illustrations, and a fully searchable ASCII text file for each page. As of February 2004, JSTOR has digitized the full back runs of more than 400 scholarly journals covering 38 disciplines and totaling over 13.4 million pages.

Successfully moving content from paper to digital media involves extensive preparatory and quality control measures that are designed to ensure that the resulting digital product will serve the scholarly community's needs for the long term. For example, JSTOR carefully assembles a complete back run and conducts a page by page review of each printed issue in order to identify any missing or damaged pages. Replacement pages or issues are located as needed in order to ensure that the complete printed copy is faithfully replicated and preserved in digital form. After the pages have been scanned by a third-party service provider, JSTOR subjects the resulting files to a rigorous quality control process. The digital files are incorporated into the JSTOR archive only after they have been evaluated and are determined to meet JSTOR's well-established quality standards.

After digitization JSTOR maintains multiple, fully synchronized copies of the complete archive both online and on various physical media. The online copies of the archive are maintained in three locations on two continents at the University of Michigan (Ann Arbor, MI), Princeton University (Princeton, NJ), and the University of Manchester (Manchester, England). Physical backup media are routinely made and are stored in New York, New Jersey, and Michigan.

The careful stewardship of print originals and the migration of content from print to digital media are important preservation components, but in some cases another component may be necessary. This is especially true for content originally made available in electronic media, such as electronic journals. Electronic journals in many ways are simply another version or edition of their print counterparts. Although there is significant overlap between the print and the electronic versions, there is often also real divergence. Over time one version--print, digitized print or electronic--may become accepted as the "copy of record"; however, until that point is reached, the safest course of action is to preserve the full range of valued journal content.

This includes finding a way to preserve the electronic journal.

**Electronic Archiving: Preserving Electronic Publications**

The challenge of ensuring long-term preservation of and ongoing access to born digital content raises significant issues not found in the world of printed publications. Electronic journals are, for example, published in more diverse formats than their print counterparts and range from fully marked up SGML/XML files, to PDF files with SGML/XML headers, to HTML files. Electronic publications may contain video or audio components or dynamic content which varies depending upon a reader's location. Functionality typical of an e-journal, such as searching or linking, adds yet another layer of complexity to preserving these materials for the long term.

In response to these complex challenges, JSTOR has launched a new effort, the Electronic-Archiving Initiative. Known more informally as "E-Archive," the initiative is focused on developing the technical and organizational infrastructure necessary to ensure the long-term preservation of and access to electronic scholarly resources. E-Archive, which is born out of JSTOR's archival commitment to the content under its care, will provide JSTOR with the capacity to process electronic journals and to add these materials to the JSTOR archive. As an operating entity separate from JSTOR, E-Archive expects to be in a position to provide long-term archival care to electronic scholarly resources which are beyond the scope of JSTOR's immediate collection development plans. In this way E-Archive's archival capacity complements JSTOR's, and working together these two entities can ensure the long-term preservation of a broader range of important scholarly resources than either entity might choose to archive alone. Over time E-Archive may have the capacity to archive a variety of electronic scholarly materials such as special image collections or digitized primary materials.

E-Archive, which is funded initially through a grant from the Andrew W. Mellon Foundation, is also being supported by Ithaka, a new organization with a mission to help accelerate the adoption of productive and efficient uses of information technology for the benefit of the worldwide higher education community. Drawing from this broad base of support from JSTOR, Ithaka, and the Foundation, E-Archive's initial activities are focused on two areas: system design and business model development. Work in these two areas, which are foundational for any trusted archive of electronic resources, is being informed by the lessons gained by JSTOR and by others. It is clear that progress can most effectively be made through collaboration with cooperating organizations that understand the importance of preserving electronic resources and that are motivated to help to develop a robust solution. In order to understand the requirements of the electronic content which an archive must preserve for the long term, E-Archive has sought the collaboration of publishers who represent various approaches to creating electronic content. These publishers have supplied sample data to E-Archive which has allowed us to pursue early analysis and development work, and they are engaging with E-Archive on a number of technical and business issues. [4]

It is also important to understand the economic implications which electronic periodicals have on library budgets, and much work has been completed in this area. In 2003, E-Archive launched a study with the support of a small number of representative JSTOR participating libraries.[5] Discussion of this study is beyond the scope of this paper, but a full report is forthcoming from the Council on Library and Information Resources and an early report is available via *DLIB Magazine*.[6] We also have reached out to numerous libraries to discuss the archiving challenge and to understand what approach will be most useful to libraries and to scholars. These activities are all contributing in key ways to the development of the electronic archiving component so necessary for long-term preservation of electronic content.

The three preservation components described above--stewardship of print originals; migration of print content to digital media; and preservation of electronic publications--can be thought of in a coordinated way. Together, they may improve the likelihood that content can be successfully preserved regardless of its original presentation medium. While each component alone provides some amount of archival protection, a greater level of protection is offered through a multi-faceted approach.

## Organizational Components

Although a multi-layered approach offers significant security, no strategy--however robust--can succeed without several organizational elements supporting and sustaining it. The 1996 Report of the Task Force on Archiving of Digital Information and the 2002 report Trusted Digital Repositories: Attributes and Responsibilities offer clear and useful descriptions of some of these necessary organizational elements. From these early reports and from the practical lessons of the JSTOR experience, it is possible to describe the critical organizational components which must be present and designed specifically to support the long-term archival objective. There are, at a minimum, five critical components that must be present in any trustworthy archive: 1) an organizational mission to fulfill the preservation role; 2) a business model to ensure the sustainability of the archival enterprise; 3) the development of a technical and content management infrastructure matched to the demands of the content to be preserved; 4) relationships with libraries; and 5) relationships with content owners and producers. Without at least these five components, the future of an archived resource cannot be assured. There may be other important components as well, but these five offer a necessary foundation.

**1. Organizational Mission**

A well-defined mission is absolutely critical because it drives the resource allocation, decision-making, and routine priorities and activities of the organization. When an organization's mission is *to be an archive* it will by necessity

dedicate its available resources to this core activity, avoiding the all too frequent competition between preservation needs and other priorities. Similarly, when long-term preservation is mission critical, preservation values and concerns will necessarily inform the shape of an organization's routine procedures and processes.

**2. Business Model**

An archive must generate a diverse revenue stream sufficient to fund the archive, including both the considerable cost of developing the archive's basic infrastructure and the ongoing operation of the archive over the long term. A single source of funding--a single donor, a single government agency, a single library, or a single foundation--should be evaluated carefully for its ability to support the longevity of the archive. Over time it is possible that noble efforts will come and go with the shifting priorities of those who control any single set of purse strings. If it is to succeed over the long term, an archive must be protected against these shifts in budgetary priorities, and a diversified revenue stream offers this necessary protection.

**3. Technological Infrastructure**

An archive's technological infrastructure must support content ingestion, verification, delivery, and multiple format migrations in accordance with accepted models such as the Open Archival Information System (OAIS) and evolving preservation practices. It must include and support the automated and manual quality control processes necessary to (1) protect the ongoing integrity of the materials, and (2) protect against format or hardware obsolescence. It also must be sophisticated enough to contend with the diversified formats in which electronic resources are and will continue to be published. Of course, no technical infrastructure lasts forever. Consequently there must be a commitment to the inevitable upgrades, rewrites and enhancements that will be necessary over the long term, and the baseline design of the technical infrastructure must facilitate these changes over time.

**4. Relationships with Libraries**

An archive must understand the needs of the library community, and the scholars whom libraries support, by building and maintaining strong relations with libraries. The archive must find a way to meet the needs of librarians and scholars in a way that balances these needs with those of other participants in the scholarly communication process.

**5. Relationships with Content Owners and Producers**

An archive must establish agreements for the secure, timely, and reliable deposit of content, and it must work with publishers and other content owners and producers to secure the rights necessary to archive the material entrusted to its care. The archive must find a way to meet its long-term obligation without impeding the ongoing work and revenue streams of content owners and producers--a difficult but important balance to strike.

These elements of a successful archive could be implemented in any number of organizational models. Indeed, the scholarly community will be best served by having multiple organizations serving as trusted archives and thereby enhancing the robustness of the community's overall archival network. But if we are to develop a network of trusted archives--and we have much work to do to reach this point--we must also ensure that each archive implements a sufficiently robust approach to preservation.

## Conclusion

Any attempt to preserve content for scholars, researchers and students must meet the fundamental preservation challenge: ensuring that content remains usable for the long term regardless of the medium which originally carried the content to the reader. Several components may be needed in order to successfully meet this preservation challenge. Those components which JSTOR's experience suggests may be helpful include stewardship of print originals; migration of print content to digital media; and preservation of electronic publications. These components, however, must ultimately be placed within an organizational context which incorporates the five elements that are critical for any long term archive. These include: 1) an institutional commitment to the preservation goal; 2) a business model which can ensure the sustainability of the preservation operation; 3) a technological infrastructure able to meet the demands of the content; 4) appropriate relationships with libraries whom in turn represent the interests of the scholars; and 5) appropriate relationships with content owners and producers.

These organizational elements together with the necessary preservation components provide a model for successfully meeting the ongoing challenge of preservation. Taken together they suggest one path toward an important goal: a trusted, reliable, and long-lived record of scholarship that meets the needs of scholars, students, faculty and researchers for generations to come.

## References

1. The terms "preservation" and "archiving" are, for purposes of this paper, used interchangeably.
2. For a detailed discussion of the value of print repositories, see Developing Print Repositories: Models for Shared Preservation and Access by Bernard F. Reilly, Jr., published by the Council on Library and Information Resources (CLIR), June 2003, www.clir.org/pubs/reports/pub117/pub117.pdf. Also relevant is The Evidence in Hand: Report of the Task Force on the Artifact in Library Collections also published by CLIR, November 2001, www.clir.org/pubs/reports/pub103/contents.html.
3. There is an ongoing debate regarding the use of emulation versus migration, the details of which are far beyond the scope of this paper. For a brief treatment of this issue see "Emulation vs. Migration: Do Users Care?" by Margaret Hedstrom and Clifford Lampe, RLG DigiNews, December 15, 2001,

www.rlg.org/preserv/diginews/diginews5-6.html#feature1.

4. We are very pleased to have the following publishers working with us on this effort: Association of Computing Machinery, American Economic Association, American Mathematical Society, American Political Science Association, Blackwell Publishing, Ltd., The Ecological Society of America, John Wiley & Sons, Inc., National Academy of Sciences, The Royal Society, and The University of Chicago Press.
5. We are grateful for support from and participation of: Bryn Mawr College, Cornell University, Drexel University, Franklin & Marshall College, George Mason University, New York University, Suffolk University, University of Pittsburgh, Western Carolina University, Williams College and Yale University.
6. See www.dlib.org/dlib/january04/schonfeld/01schonfeld.html.

## Biography

**Eileen Gifford Fenton** is Executive Director of the Electronic-Archiving Initiative. She is leading the Initiative's work to develop all of the organizational elements necessary to ensure the long-term preservation of and access to scholarly literature published in electronic form. Previously Eileen was Director of Production at JSTOR, and she has also worked at the Vanderbilt and Yale University libraries. Eileen earned a Masters of Science in Information from the University of Michigan and a Masters of Arts in English Literature from the University of Kentucky.