SPEC Kit 325

Digital Preservation

October 2011

Gail McMillan

Director of Digital Library and Archives
Virginia Tech

Matt Schultz

Collaborative Services Librarian
Educupia Institute

Katherine Skinner

Executive Director, Educupia Institute
Program Manager, MetaArchive Cooperative

ASSOCIATION OF RESEARCH LIBRARIES

Series Editor: Lee Anne George

SPEC Kits are published by the

Association of Research Libraries

21 Dupont Circle, NW, Suite 800

Washington, DC 20036-1118

P (202) 296-2296 F (202) 872-0884

<http://www.arl.org/resources/pubs/spec/>
pubs@arl.org

ISSN 0160 3582

ISBN 1-59407-869-6

978-1-59407-869-9

Copyright © 2011

This compilation is copyrighted by the Association of Research Libraries. ARL grants blanket permission to reproduce and distribute copies of this work for nonprofit, educational, or library purposes, provided that copies are distributed at or below cost and that ARL, the source, and copyright notice are included on each copy. This permission is in addition to rights of reproduction granted under Sections 107, 108, and other provisions of the US Copyright Act.

The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (R1997) Permanence of Paper for Publications and Documents in Libraries and Archives.

SURVEY RESULTS

Executive Summary	9
Survey Questions and Responses	15
Responding Institutions	75

REPRESENTATIVE DOCUMENTS

Policies, Procedures, Guidelines

University of Alberta	
Digital Preservation	80
Columbia University	
Columbia's Long-Term Digital Preservation Archive (LTA). Status Report.....	82
Preservation Asset Overview	84
University of Florida	
Florida Digital Archive.....	85
Florida Digital Archive (FDA) Policy and Procedures Guide	86
Georgia Tech	
SMARTech	91
SMARTech Mission and Collection Policy.....	93
University of Illinois at Urbana-Champaign	
IDEALS Policies and Guidelines.....	97
IDEALS Digital Preservation Policy	98
IDEALS Digital Preservation Support Policy	102
Johns Hopkins University	
What is the Data Conservancy?	107
Data Conservancy: Objectives	108
Library of Congress	
Sustainability of Digital Formats. Planning for Library of Congress Collections.....	109
University of Michigan	
What is Digital Preservation?	110

University of North Carolina	
Carolina Digital Repository. Preservation and Stewardship	112
University of Utah	
Digital Preservation. Program Overview	113
Organizational Policy Framework	114
Digital Preservation Program: Organizational Policy Framework	115
University of Washington	
ResearchWorks Archive Digital Preservation Policy	120

Cooperative Agreements

University of British Columbia	
COPPUL Private LOCKSS Network	122
University of Florida	
dLOC Digital Library of the Caribbean	128
About dLOC: Digitization	129

Job Descriptions

University of Illinois at Chicago	
Digitization Librarian & Associate Professor	132
University of Michigan	
Digital Preservation Librarian	134
Michigan State University	
Digital Curation Librarian	135
Pennsylvania State University	
Digital Library Architect	137
Digital Collections Curator	138

Data Management

Johns Hopkins University	
JHU Data Management Services	142
Michigan State University	
Lifecycle Data Management Planning	143
University of New Mexico	
Digital Data Management, Curation and Archiving	146

SELECTED RESOURCES

Books, Articles, and Reports	149
Web Resources	157

SURVEY RESULTS

EXECUTIVE SUMMARY

Introduction

ARL member libraries increasingly create, acquire, disseminate, and curate both digitized and born digital content. As a result, they have a growing awareness of and a pressing need for information on field-wide activities and plans to support the life cycle needs of these digital collections. Until now, however, relatively little information has been gathered or reported about ARL libraries' digital preservation practices and policies. This was the first SPEC survey to focus on the preservation of digital, rather than physical, materials. The definition of digital preservation includes the policies, strategies, and actions that ensure access to digital content over time.

The survey sought to identify the strategies that ARL member institutions use to protect evolving research collections and to describe the roles and responsibilities of stakeholders. It asked ARL libraries about their digital content, their strategies for preserving that content, and the staff, time, and funding they currently devote to digital preservation. It also asked each responding library to compare its digital preservation activities of three years ago to current activities and project three years into the future. In addition, to better understand the roles of research libraries in the emergent field of digital curation, the survey sought to identify issues that are and are not being addressed through current practices and policies.

The survey was conducted between March 14 and April 18, 2011. Sixty-four ARL members completed the survey for a response rate of 51%. Using the survey data and open-ended comments, this report summarizes how those libraries currently think about the preservation of their digital collections and what preservation activities they are now undertaking.

Digital Content

The survey asked what types of digital content the library licenses or manages for its institution and which content it is investing in for the purpose of preservation. Almost every library responding to this survey is responsible for managing digitized special collections, licensed materials (e.g., ejournals and databases), still images, electronic theses and dissertations (ETDs), moving images, and audio materials. Fewer than a third identified research data (including data sets and geospatial data), mass digitization collections, or art databases as current responsibilities, and only a handful manage web-harvested materials (19 or 30%) or computer games (12 or 19%).

Eighty percent of the responding libraries (51 of 64) now preserve some of their digital content and another 16% plan to do so in the future. One astute respondent commented, "This content [we have purchased or licensed from publishers] represents a significant investment of resources, whether financial, staff, or technology." Another pointed out that ensuring investments in digital preservation is "our only way to guarantee continued access to (scholarly) information in the future."

Only three respondents are not planning to preserve digital content at all. They cited the lack of experienced staff, funding for hardware and software, and institution-wide policies and strategies for digital preservation as significant barriers to preservation. Support, and ultimately approval from upper administration, for policies and strategies is deemed critical.

The categories of digital resources that most libraries are managing for their institutions are also the resources that most have chosen to preserve: digitized special collections, still images, ETDs, audio materials,

and moving images. The significance of these collections is primarily predicated on the uniqueness and overall importance of special collections and graduate student research. Respondents referred to special collections as “core (to) our identity” and “unique,” and likewise referred to ETDs as “unique output by the university community” and figure as “part of the university’s official record.”

Surprisingly, although 94% of respondents are managing licensed materials such as e-journals and databases, only 59% say that they are planning to preserve them. Ranking near the bottom of the preservation priorities are administrative records, web-harvested materials, applications/operating systems/other software, and computer games.

Local Preservation Activities

Most of the responding libraries are actively engaged in digital preservation in-house rather than outsourcing it to external parties. Ninety percent reported that they are engaged in or intend to engage in *local* activities to preserve their digital content. Half of these respondents reported that they are running digital preservation solutions in-house for their most important collections, and nearly a quarter reported that they are also running collaborative digital preservation solutions that have a local component.

Respondents described a number of factors they consider when selecting digital content for local preservation efforts. The most consistently cited criteria include local scholarly use (faculty research needs, user needs, etc.), investment level (purchased content, digitization projects, etc.), and risk factors (uniqueness, condition, etc.). Approximately 42% of respondents explicitly mentioned faculty research needs, scholarly output, and/or user needs as drivers for prioritizing content for preservation. Nearly as many give priority to content or collections that represent a significant institutional investment, including the products of digitization projects. Several institutions give priority to digital surrogates for fragile materials that preclude handling the originals. Risk factors such as uniqueness, rarity, and/or significance were also primary preservation criteria. Many institutions are making efforts to address local scholarly use, investment level, and risk factor criteria simultaneously.

When asked who will make local selection decisions, respondents most often mentioned digital initiatives librarians or collection managers. Slightly less frequently mentioned were special collections librarians and archivists. Content providers, repository managers, and library administration were least often mentioned.

Preservation Strategies: Formats

While the survey sought to gauge the current approaches research libraries use for prioritizing content and collections for long-term preservation, some of the key aims of the survey were to identify prevailing digital preservation solutions and strategies, including migration to archival formats and bit-level preservation, or combinations of these two approaches.

The question of broad support for digital formats and/or successful migration to archival quality formats has remained a topic of great interest in the digital preservation community. The survey asked if the library limits, or plans to limit, the file formats they preserve locally. Slightly more than half of the responding institutions report that they are already limiting file formats for preservation purposes. This decision is heavily influenced by concerns about format viability and technical capacity (infrastructure). As one respondent stated, “Greater uniformity of format makes management, future migration, and development of processes for ingestion, QC [quality control], and access/delivery easier.”

Several respondents mentioned the lack of available migration tools for many formats and lack of support for multiple formats in their current software repository systems. Respondents also cited a lack of financial resources as a reason for limiting formats. One institution put it succinctly, “It is ultimately an issue of time and money, in that more file types require more support. In addition, we want to focus as much as possible on archival formats (i.e., XML and non-lossy image formats) that further restricts supported file types.”

The vast majority of research libraries are committing to support only content that is deposited in an archival format or for which they have some assurances of migrating. Only a quarter of respondents have committed to more flexible support for many or

all formats, reporting either baseline bit-level preservation or some combination of format migration and bit-level preservation. Those that are not setting such limits provide credible reasons for preserving a broad range of files, including that they can preserve all formats at bit-level and consider this worthwhile for valuable resources, regardless of format. As one respondent stated, “We anticipate being able to provide bit-level preservation for any file format contributed by a member of the community that falls within the archiving scope for the repository, but will not be able to provide a full suite of preservation services for all file formats due to practical limitations such as inability to locate and implement migration tools.”

Preservation Strategies: Metadata

Fifty-one institutions reported having or creating a broad range of preservation metadata for their digital collections. Nearly all reported that they create some item-level metadata (48 or 94%), and many also create some collection-level metadata (42 or 82%). All 51 respondents reported collecting administrative metadata (e.g., access privileges, rights, ownership of material), and all but one also collect technical metadata (e.g., information describing the production process or digital attributes of the work). Slightly fewer (approximately 84%) report collecting metadata about structure or provenance at this time.

Fifty respondents reported using multiple schemas to describe their digital collections. Of these, the most popular metadata formats are Dublin Core (40 or 80%), Qualified Dublin Core (35 or 70%), and METS (35 or 70%). Slightly more than half (26 or 52%) also reported using PREMIS. As is typical in the ARL community, many reported using additional metadata schemas in their digital collections management practices, including EAD, NLM, FGDC, IPTC, MIX, TEI, RDF, MARC, VRA Core, PBCore, AESS, and Darwin Core.

Preservation Strategies: Policies

The survey sought to gauge progress toward the development and adoption of formal digital preservation policies that have been well researched in regard to prevailing standards, are developed with key stakeholders, and have a goal of securing support from upper administration.

Collaboration is a significant factor in current preservation planning and activities. A solid majority of respondents (42 or 70%) are working with other stakeholders within their parent institutions as they make decisions about digital preservation policies and investments. Most of these are working with campus IT, faculty, and administration.

Policy development is underway in a large majority of the responding libraries, but only two institutions have approved digital preservation policies in place. Discussion of preservation policies is underway at 27 of the responding libraries (44%), and 13 (21%) have written drafts. Of those libraries that are in discussion and draft stages, the majority are approaching policy development as a campus-wide initiative, inclusive of stakeholders beyond the library such as campus IT, university archives, offices of scholarly communication, offices of strategic initiatives, and digital services, among others.

On the whole, the responding libraries are consulting well-developed, community-derived digital preservation standards. These include resources such as the Reference Model for an Open Archival Information System (OAIS), the Trustworthy Repositories Audit & Certification: Criteria & Checklist (TRAC), JISC's Digital Preservation Policies Studies, along with the Interuniversity Consortium for Political and Social Research (ICPSR) and Cornell's Digital Preservation Policy Framework, among others.

Based on respondents' comments, it is much more likely that a group within the library, rather than an individual, will have primary responsibility for researching and developing the library's digital preservation policies. These groups are not likely to have membership from outside the library. In the relatively few libraries that give an individual policy development responsibility, it is typically a digital initiatives librarian or special collections head.

Similarly, the authority to approve the library's digital preservation policies and investments resides with a library group, which usually includes a library administrative team. A majority of respondents (60%) indicated that library administration has primary responsibility for authorizing and approving digital preservation policies. Only a few explicitly indicated that an authority external to the library (e.g.,

the university president, vice provost, university IT, or campus CIO) would have a role in approving the library's digital preservation policy.

Resources and Funding

Most of the respondents report they are now funding digital preservation through a mixed revenue model that includes a range of internal and external funding sources. The good news is that 83% of respondents report that their libraries fund at least part of their digital preservation activities through their general operating budgets. More than a third report having a dedicated preservation budget. Many also report that other internal funding lines, including their IT budgets (62%) and their materials budgets (38%), cover a portion of their digital preservation work. Grants and awards still provide a hefty percentage of funding (38%), and some institutions (35%) report even having gifts and endowments as an additional, and growing, funding source for digital preservation. Almost all expect their funding to increase or at least stay about the same in the next three years. Interestingly, only two respondents speculated that funding might decrease in part because "...there will no longer be the costs of setting up various parts of the preservation activities."

Survey respondents' comments reveal that funding fluctuations, both positive and negative, are often tied to grant money, including state funds, National Science Foundation (NSF) grants, and National Digital Information Infrastructure and Preservation Program (NDIIPP) awards. Other respondents referred to the shift from print-based work to digital work and the resulting increase in funds available for digital preservation, though as one respondent noted, "The shift is slow."

When asked to compare today's levels of investment in staff, time, and funding to the investment levels of their libraries three years ago, the majority of respondents reported that they are investing more. Two-thirds say they have more staff devoted to digital preservation, three-fourths say they are investing more time, and 60% say that they are spending more money on digital preservation. Only three respondents (6%) report that they are investing less staff and time, and seven (15%) are investing fewer dollars in digital preservation.

Twenty-nine of the 45 university libraries (64%) have from one to three FTE responsible for digital preservation. But at seven libraries there is less than one FTE. Usually the digital preservation responsibilities are divided among two or more library staff and only rarely is an entire FTE embodied in one individual.

Barriers to Digital Preservation

The survey sought to gauge both the willingness and capacity of respondents to keep pace with the growth of digital content at their libraries. Not surprisingly, almost all of the respondents (46 or 94%) stated that their libraries want to invest in preserving more digital content than they currently do, but their comments indicate they face a number of similar barriers to additional efforts. The most frequently reported barriers to preservation were staffing and expertise. The responding libraries are struggling to dedicate staff to digital preservation and to foster staff expertise to keep pace with the technical challenges inherent in digital content, technical infrastructures, and digital preservation best practices.

Funding and resources for technical development, equipment purchases, and support for on-going operations were also frequently cited barriers. Several institutions reported having difficulty making the transition from grant-funded support to dedicated institutional funding for sustained operations. Finally, several libraries reported that their institutions lack clear institutional policies and/or strategies for guiding investments. Other less-cited, but still significant barriers include legal issues regarding deposit, lack of trustworthy repository status, and the absence of reliable standards for complex digital data.

Thirty-four libraries reported they plan to manage a digital archive/repository that is intended to support preservation functions. But strikingly, 70% of these respondents reported that some institutional units, including academic units, administrative units, and data centers, are "indifferent" to deposit, or are "not actively seeking deposit." They cited several commonly perceived and expressed barriers to deposit, such as awareness, library capacity (real or perceived), complicated submission workflows, and concerns about future access to their content.

Despite real and perceived barriers to digital preservation progress, the responding libraries are moving forward—planning to preserve more content, improve their strategies, and develop policies that will better ensure the long-term viability of their digital assets.

Future Digital Content

Only a few of the libraries responding to the survey are satisfied with maintaining their current levels of digital preservation. As mentioned previously, most libraries do want to preserve more, especially collections such as research data, geospatial data, various media, faculty research, university history, and web content. Those that are not already hosting and preserving ETDs and digitized special collections commented that these would probably receive their attention in the future when additional resources allow expansion of their activities. Most reported conditions similar to one respondent who stated, “All areas will require more, set by collection priorities and risk.”

Future Preservation Strategies

Compared to three years ago, most of the responding libraries are currently investing more staff, time, and funding in their digital preservation activities. The majority anticipate that this trend will continue over the next three years. As one respondent said, “Increased reliance on digital resources has made this imperative.” Another commented, “As the library’s digital collections grow in size and diversity, so too will the need for staff working in all aspects of digital preservation.” Only four libraries expect their investment to decrease. As one respondent stated, “It’s hard to predict an increase in our funding/budget situation going forward given the current climate. As a result, we can only be pessimistic for purposes of this survey and expect the worse: further budget cuts or at best, level funding. We do continue to actively pursue research grant opportunities, however.”

Currently, respondents use a range of strategies for preserving most digital resources. Their first choice solution is using a library-managed digital archive/repository. This strategy is followed by collaborative solutions, either with other administrative and/or technical units in the institution, in a participatory

solution such as the MetaArchive, or in a hosted solution such as the HathiTrust.

While the majority of respondents predict that using a library-managed digital archive/repository will remain their primary strategy, an increasing number anticipate that participating in collaborative solutions will be part of their future strategy. Nearly 25% of those that expect to collaborate are not currently collaborating as part of their preservation strategy. A vendor-based solution is the least likely future preservation strategy. Among the “other” anticipated strategies, respondents mentioned homegrown solutions and institutional and statewide repositories.

When attempting to explain why future strategies might be different from their current preservation strategies, three reasons were cited most frequently: 1) They are not now, but they plan to collaborate. 2) Their repositories will develop further. 3) They will take advantage of third-party or remotely hosted solutions (HathiTrust usually). Also mentioned, but less often, were changes due to centralization of efforts within their institutions.

Training

Research libraries are turning to institutional peer staff and seeking broader community-based opportunities to improve expertise in digital preservation. The vast majority of respondents reported that conferences and workshops are the primary methods used to increase staff expertise. Independent study is another frequently used method. Thirty-six respondents (62%) take advantage of training provided by professional organizations. Fewer look externally to vendors or consultants. Several rely on in-house training or presentations by library staff.

When asked what types of services their library would find valuable for improving its role in preserving digital content, respondents identified standards/best practices (81%), preservation planning (76%), and policy recommendations (75%) as their top needs. A slightly smaller majority expressed a need for technical training (71%) and conversion/migration services (61%). Interestingly, slightly less than half would find appraisal and selection training valuable, and one-third want theory training. This may indicate that research libraries are eager to move past conceptual

decision-making and are beginning to make practical progress in accomplishing digital preservation.

Conclusion

ARL libraries curate a diverse and growing range of digital collections that include digitized and born-digital special collections, licensed materials (e.g., ejournals and databases), research data, art databases, web-harvested materials, administrative records, and electronic theses and dissertations (ETDs). The curatorial challenges they face for these assets are acute. The collections often began with ad-hoc and idiosyncratic data storage structures resulting from project-driven needs (e.g., to host scanned copies, to amalgamate data in a variety of formats and databases, or to establish an effective workflow for accepting born-digital works). Of necessity, the libraries have allowed these collections to expand and have regularly acquired new digital collections over the last several decades before they could implement clear mechanisms for the preservation of this digital content.

Today, methods for preserving digital content are becoming standardized and digital preservation models (e.g., MetaArchive, UC3 Merritt, DAITSS, HathiTrust) are readily available in the field. This survey revealed, as the digital preservation field is maturing, that most ARL libraries are rising to the challenge of establishing policies, workflows, and infrastructures to systematically preserve their rapidly expanding bodies of digital content. The survey also revealed that most ARL libraries are actively engaging in in-house digital preservation rather than outsourcing it to external parties, thus maintaining their control and ownership over the digital content that they curate. Survey respondents also predicted that they would continue turning to library-managed and collaborative solutions over vendor-based, hosted solutions for their core collections.

Tempering our excitement at the unprecedented levels of reported preservation activity are some of the comments made throughout the survey that demonstrate that the definition of “digital preservation” is still murky for some librarians. A number of respondents confused “back ups” with “preservation” and referred to access-oriented repository services as though they were preservation solutions. For

example, respondents stated that they are “organizing and backing up digital assets in-house,” and named non-preservation services, such as Archive-It, as their preservation strategies. However, others are quite sophisticated in their understanding of preservation and their responsiveness to the current environment, including one member who reported, “We’re keeping our eye open for the most effective strategy...right now it is hedging by employing multiple options.” This mixture of responses demonstrates that there is still a serious need for training opportunities in digital preservation and life-cycle curation for the ARL community.

Judging by the survey findings, most ARL libraries view digital preservation as a complicated mix of technical and organizational responses to the needs of aging content. Most also see the provision of digital preservation services for their campuses as a key component of their 21st century missions. They are actively expanding their policies, workflows, and technical capacity for preservation.

This expansion is, in itself, challenging. It requires a paradigm shift in thinking about the library’s mission as an active caretaker of non-physical content; it also requires heavy resource allocations to establish a solid infrastructure for digital life-cycle curation. However, there is a second challenge that ARL libraries cite and must respond to at the campus level. Respondents report that other campus entities (e.g., research data centers, administrative units) are often both unaware of the library’s growing capacity for digital curation and ambivalent at best about engaging the library’s services for their own data collections. If ARL libraries are to maintain their core role as the campus’s source for collecting, providing access to, and preserving not just analog but also *digital* collections, they must find new ways of engaging with their campus constituents, including through advertising these services and engaging directly with the content producers. Doing so will help to ensure that the campus turns to a central entity—the library—to maintain its scholarly communications channels and materials in the increasingly digital age, rather than distributing this responsibility across other campus units or outsourcing it altogether.