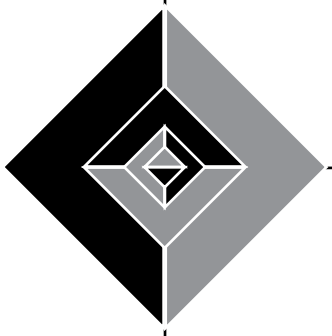


SPEC Kit 329

Managing Born-Digital Special Collections and Archival Materials

August 2012

Naomi L. Nelson and Seth Shaw

Duke University

Nancy Deromedi and Michael Shallcross

University of Michigan

Cynthia Ghering and Lisa Schmidt

Michigan State University

Michelle Belden, Jackie R. Esposito, Ben Goldman, and Tim Pyatt

Pennsylvania State University

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 1-59407-883-1 / 978-1-59407-883-5 print
ISBN 1-59407-884-X / 978-1-59407-884-2 online

SPEC Kit 329

Managing Born-Digital Special Collections and Archival Materials

August 2012

SURVEY RESULTS

REPRESENTATIVE DOCUMENTS

Job Descriptions

Collection Policies

Gift/Purchase Agreements

Format Policies

Workflows

SELECTED RESOURCES

SURVEY RESULTS

EXECUTIVE SUMMARY

Introduction

The 2010 OCLC Research report, *Taking Our Pulse*, listed management of born-digital materials as the third biggest challenge facing libraries, special collections, and archives, after space and facilities. It has become a truism that the trickle of born-digital materials into special collections has become a flood. Increasingly, these materials do not have analog counterparts. Libraries and archives can no longer defer decisions about digital content to a later date. We must develop policies and procedures to operationalize the management of born-digital materials, or we risk losing the record of the recent past.

This survey sought to gather and promote emerging good practices for managing born-digital content and to highlight common challenges. The survey instrument focused in particular on staffing, ingest and processing workflows, storage procedures, and access and discovery methods. Sixty-four of the 126 ARL member libraries responded to the survey between February 22 and March 23 for a response rate of 51%. Fifty-nine of the respondents (92%) already collect born-digital content. The remaining five libraries are in the planning stages. The level of engagement with born-digital content was higher than anticipated by the survey team. An analysis of the responding libraries engaged with born-digital materials revealed they are larger institutions and therefore more likely to be pioneers in working with this content.

The management of born-digital materials is still relatively new for ARL libraries, and the survey results show that good practices and workflows are still evolving. New tools are emerging rapidly, and the once-solid line between digitized content and born-digital content is beginning to blur. Survey responses indicated that the library and archives profession lacks a common definition of what born-digital content is and a common understanding of who within the organization should manage this content.

Staffing and Organization

The survey asked how many library staff collect and manage born-digital materials, who has responsibility for storage-related activities, how staffing needs are addressed, and how staff gain the expertise required to manage these materials. No one staffing or organizational structure emerged from the survey responses, which again reflects the evolutionary status of born-digital management programs.

The number of staff working with born-digital archival content in the responding libraries ranges from less than one to 60 FTE. While archivists and librarians in institutional and government archives were the trailblazers in collecting this content, managing these materials now requires staff from digitization, digital curation, information technology, and institutional repository units. Respondents most frequently mentioned special collections/archives staff and library IT staff as having decision-making responsibility for selecting storage solutions, implementing and maintaining infrastructure, managing user authentication, estimating storage needs and monitoring usage, and budgeting. Many other units are also involved, including institutional IT, preservation, collections, administration, and consortia in a wide variety of combinations.

This organizational distribution may factor into how respondents have addressed staffing needs for

managing born-digital content. Almost all have used a combination of strategies, either adding that responsibility to existing positions (94%) or recasting an existing position (37%), and creating new positions (46%). Training strategies reflect the emphasis on retooling the skill sets of existing positions. Conferences, on-the-job training, workshops, and independent study are the primary methods staff use to develop their expertise with born-digital content.

Born-Digital Materials Collected

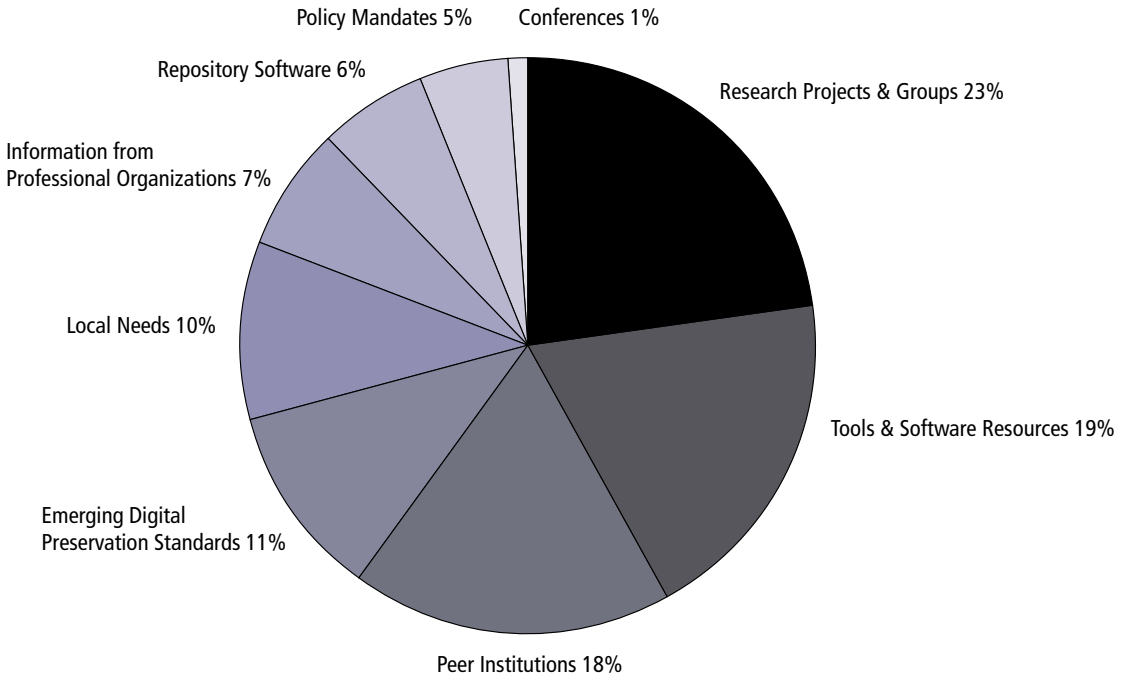
Almost all of the responding libraries (54 or 84%) are currently collecting electronic theses and dissertations. The majority also collect personal archives and institutional records and archives. Most of the others report they plan to collect these categories of materials. Twenty-one libraries collect research data and 28 others plan to collect it. Photographs, audio and video recordings, texts, and moving images are the most frequently collected media formats. About a third of the respondents collect websites, email, and databases; almost an equal number plan to collect these formats. While only six currently collect social media, 23 others plan to do so in the future.

Ingest Policies and Procedures

The majority of respondents (45 or 71%) have not developed gift/purchase agreement language that is specific to born-digital materials, but many are reviewing those agreements. Thirty-six respondents (56%) reported that they have developed ingest and processing workflows. An analysis of the comments indicates that a number of libraries are in the development phase. The comments also revealed a variety of models and/or examples the libraries have used in the development of workflows. These influences can be grouped into nine general categories as seen in the chart below.

Projects that influenced workflow development include the Personal Archives Accessible in Digital Media (PARADIGM) and futureArch projects at the University of Oxford's Bodleian Library, the AIMS project (Born Digital Collections: An Inter-Institutional Model for Stewardship) conducted by Stanford University, Yale University, University of Virginia, and University of Hull (UK), InterPARES, the British Library's Digital Lives project, the Tufts Accessioning Program for Electronic Records (TAPER) project, the European Union's Preservation

Influences on the Development of Ingest and Processing Workflows



and Long-term Access through Networked Services (PLANETS) project, and the Sustainable Archives & Leveraging Technologies (SALT) research group at the University of North Carolina.

Influential tools and software resources include Archivematica, the Duke Data Accessioner, digital forensics tools (including AccessData FTK Imager), file identification and validation tools (such as DROID and JHOVE), and the University of North Carolina's Curator's Workbench.

Respondents highlighted documentation made available by the Interuniversity Consortium for Political and Social Research (ICPSR) at the University of Michigan, the Digital Preservation Management workshop developed at Cornell University, the University of Illinois's IDEALS (Illinois Digital Environment for Access to Learning and Scholarship) repository, the California Digital Library's Merritt repository, Stanford's digital forensics lab, Emory University's Salman Rushdie collection, and Chris Prom's Practical E-Records blog.

Standards that influenced workflow development include the Open Archival Information System (OAIS) Reference Model, the PREMIS (PREservation Metadata: Implementation Strategies) metadata schema, the SWORD (Simple Web-service Offering Repository Deposit) protocol, and the BagIt specification.

Information provided by the MetaArchive, the National Digital Information Infrastructure and Preservation Program (NDIIPP), and professional journals, as well as the Digital Curation Centre's lifecycle model, influenced several respondents.

Perhaps as a sign of how workflows are tailored to fit local resources, some respondents cited DSpace repository software and CONTENTdm as influences on workflows. A few cited policy guidelines and mandates from parent organizations. Others mentioned Society of American Archivists and Midwest Archives Conference panel presentations on practical approaches to born-digital records, although no one mentioned conferences such as iPRES or the Personal Digital Archiving conference for which born-digital content is the specific focus.

While it appears that many respondents do not yet have well-established workflows for the ingest and processing of digital content, the majority are actively addressing the challenges of preparing born-digital content for long-term preservation and access.

Ingest Strategies

Seventy-seven percent of the responding libraries are ingesting born-digital records that are stored on legacy media. Almost all of them are storing the media "as is," and about half are collecting hardware that can retrieve data from those media. Fifteen libraries (25%) are outsourcing data retrieval and another 20 (33%) are planning to use that strategy. Only eight libraries are building new systems that replicate the function of the legacy systems. Other strategies include migrating content from legacy media to a storage location (described variously as "server storage" or "dark archives" space) and converting legacy born-digital content to "modern," "less proprietary," or "the latest usable" formats that include CSV files and PDF/A files.

Storage Solutions

The survey asked which kinds of storage media are used for ingest, processing, access, back up, and long-term dark storage functions. Most respondents use a combination of external media, network file systems, and local storage for all functions. Only 12 respondents (19%) report using cloud storage.

Local/attached storage (46 responses or 75%) and external media library (41 or 67%) were the most prevalent ingest solutions, followed closely by a network file system (35 or 57%). Other solutions include the DSpace-based commercial hosted Open Repository, the OnBase commercial enterprise content management system, and an institution's collection development instance of DSpace. One respondent stated that they are currently using cloud storage on a limited basis for ingest, and "plan to investigate its use for the other categories." Another belongs to a consortium that provides web-based ingest, processing, and access for ETDs, presumably including storage.

The most prevalent processing storage solutions are a network file system and local/attached storage, both at 43 responses (75%). External media library was a distant third. Other solutions were the same as for ingest: the consortium, the collection development instance of DSpace, and OnBase.

The most used access storage solution is a network file system (43 responses or 72%). External media library and local/attached storage each received 27 responses (45%). One respondent noted that they use Amazon Cloud and hosted Open Repository. Another uses a local DSpace instance, the California Digital Library's web archiving service, and a university system-wide open access repository. Other solutions include the use of a local implementation of a Fedora repository, YouSendIt online file sharing software in combination with e-mail, and shared IT servers.

The most common back up storage solution is a network file system (44 responses or 76%), followed by external media library (31 or 53%), local/attached storage (23 or 40%), and distributed systems (16 or 28%). Other solutions include a combination of Amazon Cloud and hosted Open Repository, the California Digital Library's Merritt Repository, redundant storage managed by campus and library IT, and physical tape storage.

Network file systems are used most for dark storage (26 responses or 52%), with distributed computing/storage systems second (19 or 38%). External media library and local/attached storage were not far behind at 16 and 14 responses, respectively. Other dark storage solutions include the California Digital Library's Merritt Repository, the Chronopolis digital preservation network, the Isilon commercial storage platform, redundant storage managed by campus and library IT, and virtual and physical tape storage. One respondent stated that rather than dark storage, their institution uses Fedora as an asset management system and copies files to "replicated storage for long-term preservation, with appropriate preservation metadata and restricted access."

Estimating Storage Needs and Costs

Twenty-six of the responding libraries (59%) estimate future digital storage needs and costs based on past and current usage and/or planned growth. Three noted that storage is allocated on a case-by-case basis. Some respondents have yet to implement methods of estimating storage needs and costs. Others are in the process of developing such methods.

Respondents described a variety of approaches to estimating storage needs and costs. One is conducting a longitudinal analysis of trends in digital storage growth. Another will scale future digital storage needs to the "development of campus department operations." Another currently uses costs of disks, storage devices, and backups as the basis for total cost estimates and is looking at moving to endowment-based storage cost models in the future. One respondent anticipates using the L.I.F.E. (Life Cycle Information for E-Literature) model developed by University College London (UCL) and the British Library for estimating curation costs, including the cost of storage.

One institution estimates space needs based on "past collecting volume + a 20% inflator + any known collections we anticipate receiving." Another estimates required storage needs based on average file size for a particular type of record and then estimates costs based on the current market value of storage, "usually at the TB level."

The most detailed response described the institution's attempt to estimate storage needs by tracking historical usage and growth, contrasting those with earlier projections, and categorizing data by type to identify growth areas. Thus far, the respondent observes that "consumption generally increases by a factor of 2 to 4 within a 12–18 month period," but any projection can change when unexpected projects or changes in the organization occur.

Access and Discovery

The survey asked which delivery methods the library uses to provide access to born-digital materials. Two-thirds of respondents provide online access to a digital repository system. Just under half provide in-library access on a dedicated workstation. Users who bring their PCs to 22 of the responding libraries can access born-digital materials stored on portable media. Eighteen respondents (28%) use third-party systems such as CONTENTdm, Archive-It, Dropbox, and YouTube to share materials with researchers.

There is not one, single repository system being used either to manage or provide access to born-digital materials. Most respondents use open source repository software for both management and access functions. Twenty-eight institutions report using secure file system storage to manage collections but only

ten use it to provide access. The results seem to suggest that access to collections is not as fully developed as the management of born-digital content.

The survey asked whether the institution is using different types of repositories for different types of born-digital materials. While 63% reported that they are, their comments indicate that they use different repositories for a variety of reasons, including media type (e.g., images, audio/visual materials, websites), record type (e.g., thesis and dissertations, faculty preprints), access and preservation requirements, and whether the material is digitized or born digital.

Ingest Challenges

The challenges related to the ingest of born-digital materials can be grouped into three broad categories: the difficulties associated with accessing information stored on legacy media and/or in obsolete file formats; the lack of policies, end-to-end workflows, and robust, integrated systems for digital object ingest; and the need to scale up to meet the increasing volume of born-digital objects needing preservation.

The challenges related to working with legacy formats and hardware were the most frequently cited ingest issues (43% of respondents listed file format or software obsolescence; 38% included legacy media or hardware). Donors, campus offices, and other records creators place their materials in a library or archives when they are no longer actively using them. As a result, libraries often receive storage media (punch cards, floppy disks, hard drives, CDs, zip disks, etc.) that are no longer accessible through current technologies.

Being able to transfer the files to appropriate storage is only the first step. The archivist then needs to be able to open them to assess their content. Obsolete file formats sometimes cannot be opened or executed using current software. Older versions capable of opening the files might require specific environments (operating systems and hardware) to run. Copyright restrictions and the terms of software licenses may make it difficult or impossible for staff to locate versions they can legally use. In addition, digital objects accessed through more modern systems often render differently than they did in their original environment. The formatting or appearance may be altered, and sometimes the behavior or even the actual content will change. Without the ability to access the content of older digital objects, it is difficult to determine which digital materials are most important and how best to allocate resources among collections. Given these challenges, nearly three quarters of respondents reported that their institutions store at least some of their legacy media as is, without transferring to new media or to server storage.

Collection donors have used a very wide variety of hardware and software configurations over time. As one respondent noted, "Each new collection seems to bring new technical issues that must be dealt with." In most libraries, it is unclear who should be responsible for developing technical solutions for accessing legacy media and obsolete file formats. This work is often outside the mandate of the information technology division and usually beyond the expertise of special collections staff. Some libraries and archives are creating "ingest labs" in house (the Bodleian Library, the British Library, Stanford, and the University of Virginia have working labs that serve as potential models). Others are outsourcing file recovery. An alternative file management strategy is to use a tool such as the Catweasel universal floppy disk controller, which is designed to connect legacy floppy disk drives to modern computer systems so that data can be read and written to floppy disks.

Interestingly, few respondents discussed challenges associated with complex digital objects (comprising more than one file and/or more than one file type), social media, digital objects stored in the cloud, websites, and networks of information, presumably, because most special collections and archives are just beginning to work with these types of digital objects.

The second category of ingest challenges relates to the workflows and systems needed to manage the digital objects once they are transferred off of their original carrier media. Maintaining privacy and providing adequate security topped the list of concerns. Respondents called for privacy and security policies specific to digital objects that address donor concerns and that insure compliance with university policies and federal and state laws. They noted the need for secure storage and networking and for tightly controlled access to files that contain personally identifiable

information. (See Kirschenbaum, *Digital Forensics*, pages 49–58 for additional discussion of privacy and security issues related to born-digital objects.)

Several respondents noted that archivists need to be able to dedicate more time to developing policies and conducting test pilots. The lack of clear policies and workflows can lead to inconsistent practices across collections and across the institution, and to inefficient resource allocation. Without consistent policies and procedures libraries cannot insure continued access to the born-digital objects. The PARADIGM project (Bodleian Library) and AIMS project both provide guidance in establishing policies and workflows. The BitCurator Project, led by the School of Information Science at the University of North Carolina at Chapel Hill and by the Maryland Institute for Technology in the Humanities at the University of Maryland, is building on these efforts. It will define and test a digital curation workflow, beginning at the point of encountering holdings that reside on removable media and ending with interaction with an end user.

The tools and systems used in the ingest process tend to be modular, and many were originally developed for use by other communities. For example, commercial forensics packages (which are very useful for browsing content and identifying personally identifiable information) were developed specifically for law enforcement. While the functionalities of these products have guided institutions in the development of workflows, they cannot be easily combined to meet the needs of the library and archives community. As one respondent noted, "There are several open-source and commercial products that can do pieces of the workflow, but as they are not designed to work together there are inefficiencies in stringing these workflows together." Another respondent added that "most ingest software is in alpha or beta release, with long-term roadmaps for future development." Early adopters and those libraries able to develop their own systems need to be comfortable with uncertainty and a certain amount of churn. Other archives are waiting for system development to catch up with their needs. Systems currently used include Archivematica, Rosetta, and the Curator's Workbench; others like Hypatia and BitCurator show potential for the future.

The final category of challenges related to ingest relates to the capacity needed to scale up workflows and systems to manage the flood of born-digital objects needing preservation. Respondents highlighted the need for sufficient storage space, adequate network capacity, increased staffing, staff training, automation of standard tasks, and enterprise-level systems. One respondent noted, "Our current archival storage was scaled to accommodate our analog to digital digitization program." It is more challenging to estimate the needs for born-digital special collections and archival materials: the timing for acquisitions can be hard to predict; the volume is not always known at the time of receipt (often because the digital objects are on legacy media); the formats often vary widely; and it is often unclear which materials will need to be restricted (because the files cannot be accessed before receipt due to media or format).

Storage Challenges

The challenges related to storage systems can be separated into three major areas: systems limitations, organizational challenges, and insufficient resources (i.e., not enough available space and high storage costs). The challenges surrounding systems limitations were divided between the need for preservation-quality infrastructure and the need for security for and access to the materials themselves. Organizational challenges fell into three categories: policy and planning, gaining and retaining sufficient staff and skills, and managing the organizational structure (from the department up to the entire organization) while maintaining effective coordination between all the stakeholders. One set of concerns about sufficient resources represents two sides of the same coin: insuring adequate file storage space and its cost. Other challenges related to storage space include the difficulty in estimating and predicting capacity needs. One comment that summarizes the issues well indicates that storage needs for born-digital records should not be only the responsibility of the library and archives: "Future storage needs for large-scale ingest of born-digital special collections materials will probably be integrated into university-wide planning for digital repositories, a digital asset management system, and networked storage and continuity planning."

Access Challenges

The biggest access and discovery challenge, described by 32 respondents, is the sensitivity of materials—concerns about copyright, confidentiality, privacy, intellectual property, and personally identifiable information. The second biggest challenge is IT infrastructure, or rather, the lack of it (28 respondents). Particular concerns in this area include user interface, the need to integrate multiple systems, and the ability to handle very large files. Other significant challenges are the need to develop policies, processes, and tools for arranging and describing born-digital materials in ways that make them most accessible, including the integration of description for digital and non-digital materials; rights management (restrictions specific to users rather than materials); and staff time and skills. Interestingly, *time* was twice as much of a concern for respondents as *staff skills*. This makes sense as more professionals are assigned responsibility for these materials and go on to develop the necessary skills, but *staff* may still mean the only person, or one of a very few, responsible for managing these types of materials at their institutions. The remaining concerns included metadata standardization, differing levels of donor restrictions and how to apply them in an online environment, format standardization and migration, and institutional support (including funding).

Respondents' concerns grow even more complex when restrictions on sensitive materials (those subject to copyright, confidentiality, privacy, and intellectual property concerns) are combined with rights management by user group and donor-imposed limitations on access, because each of these types of restrictions can vary from case to case. Reference desk staff have dealt with the complexity of access restrictions in face-to-face transactions for decades, but libraries lack automated systems that can do the same during online transactions where staff are not there to intervene.

Respondents' comments on registration procedures highlight the nature of this challenge. Most institutions that provide access to born-digital materials are either doing so in their reading rooms and following standard reading room registration procedures or are providing access to the materials online with no registration procedure. These limited approaches are directly linked to the second biggest access challenge for respondents, the lack of a fully developed IT infrastructure for delivering born-digital materials to researchers. Other technology concerns include user interface design, the need to navigate multiple disconnected systems, and problems supporting large file sizes.

Providing access to archival materials is, of course, dependent on appropriate arrangement and description, and so it should be no surprise that many respondents stated a need to further develop policies, processes, and tools for arranging and describing born-digital materials in ways that make them most accessible, including the integration of description for born-digital, digitized, and non-digital materials.

The survey results indicate that our profession is moving towards a higher comfort level with the standardization of both metadata and file formats. Furthermore, institutional support is a challenge at only three institutions, which would seem to illustrate administrators' growing understanding of the need to support access to born-digital materials. Possible areas for future research include the use of analytics and user studies to track the quantitative and qualitative aspects of access to these materials by off-site researchers and the challenges of providing not just basic access but value-added reference services to those users.

Privacy Concerns

The survey team was surprised that most respondents did not address the potential institutional liability posed by personally identifiable information (PII) within born-digital materials, beyond the imposition of access restrictions. (PII includes information such as social security numbers, credit card numbers, logins, passwords, PINs, and medical and financial records.) Seventy-one percent of respondents indicated that their gift agreements did not include language that acknowledged born-digital materials. While ownership transfer, copyright, and some standard restrictions can be handled through the traditional deed of gift, gaining permission from the donor to use forensic tools that allow recovery and review of deleted files while searching for PII is not a standard option. Since such actions might alter donated files or uncover files not intended for transfer, requesting permission through

the gift agreement or some other policy document is highly recommended.

While 71% of respondents have policies regarding whether files with PII should be retained with restrictions or destroyed, only 43% have policies indicating whether born-digital materials can be made available for research use before they are screened for PII. One respondent's comment that "all special collections materials have personally identifiable information (PII)" is quite true. However, paper-based collections have always benefited from security through obscurity. There is no fast or easy way to uncover social security and credit card numbers in paper-based collections. With born-digital records, on the other hand, there are many tools available that can search and locate PII, even in deleted or hidden files. Such content, improperly managed, not only puts the file creator at risk, but also may be in violation of an institution's security and privacy policy for this type of information. Eighty percent of respondents indicated that they do not have a written PII policy. Greater security is needed for unscreened born-digital records, especially if they are stored on networked servers.

Conclusion

The responses to this survey indicate that many ARL libraries and archives have begun working with born-digital materials in their collections, despite the fact that enterprise level systems and best practices for managing these materials in an archival setting are still in development, and despite concerns that they do not have the resources to scale their work to meet current and future demand. This willingness to experiment, to learn new skills, and to seek to understand the scope of the issues is building expertise within the library and archives profession, and has insured access to some born-digital holdings, at least in the near term. It also signals a shift from a wait-and-see attitude to a more empowered something-is-better-than-nothing approach to managing born-digital materials.

Respondents identified the following as critical for transitioning their work with born-digital materials from projects to programs:

- Collaborative solutions for dealing with hardware and software obsolescence.
- More, and more appropriate, storage for born-digital materials (long-term, authenticated, secure, verified, backed-up, and geographically distributed). As one respondent noted, "Archives are guaranteed preservation only if stored on enterprise data storage."
- Automation of as much of the workflow as possible.
- Asset-level access control to enable tiered access to restricted records.

Many institutions are working with digitized content or licensed digital content and are only now beginning to explore the ways in which born-digital, primary-source materials may be different. For example, it is difficult to estimate storage needs for born-digital primary sources stored on legacy media prior to accessioning and processing them. Privacy concerns are magnified when large bodies of easily searchable digital material may contain personally identifiable information. The workflows and infrastructure built for digitized content are often insufficient for born-digital primary sources.

While some special collections rely on a single staff member to manage all aspects of preserving and providing access to born-digital materials, more frequently staff from special collections, library IT, digital repositories, digital curation, and other areas work together to ingest, appraise, describe, preserve, and provide access to this content. The distributed nature of this model allows the library to leverage existing expertise, but it may also mean that no one has the big picture. These situtions make it difficult to track the resources needed to manage the materials—which then makes it difficult to estimate current and future costs. Distributed responsibility can also threaten the long-term survival of the materials, either when no one feels empowered to make decisions or when someone makes decisions without having all of the relevant information. Staff need models of existing

teams that describe how responsibilities are assigned and decisions are made collaboratively.

Survey responses indicate that best practices will take some time to develop: infrastructure, systems, and tools are in development; libraries continue to experiment with organizational models to find those that will be most effective to manage born-digital, primary-source materials; and the variety of record formats continues to grow. While several libraries and archives have developed scalable solutions that work within their own context, few of the solutions developed to date have been transferable between institutions. Waiting for time-tested systems and practices, however, is not an option. For now we need to settle for "good enough" practice and continue to invest time and resources in developing systems and workflows that will prevent a "digital dark age" for the first part of the 21st century.