# NATIONAL AGENDA FOR DIGITAL STEWARDSHIP

2015


NDSA

# Executive Summary

The *National Agenda for Digital Stewardship* provides funders, decision-makers, and practitioners with insight into emerging technological trends, gaps in digital stewardship capacity, and key areas for research and development to support the work needed to ensure that today's valuable digital content remains accessible, useful, and comprehensible in the future, supporting a thriving economy, a robust democracy, and a rich cultural heritage.

This *Agenda* integrates, annually, the perspective of dozens of experts and hundreds of institutions. The *Agenda* is released by the *National Digital Stewardship Alliance*, a membership organization of leading government, academic, nonprofit and private sector organizations with digital stewardship responsibilities. Members of the NDSA collaborate to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations.

The *Agenda* outlines the challenges and opportunities related to digital preservation activities in four broad areas: Key Issues in Digital Collection Building, Organizational Policies and Practices; Technical Infrastructure Development; and Research Priorities. Each section articulates priority challenges, and then offers a set of *Actionable Recommendations* to address the challenges.

**Changes in the Climate for Stewardship**

The last ten years have seen strong global trends in the production and use of digital content. The theme of the decade has been *more*: more information being produced; more content being published and shared; more forms of publication and filtering; more public access to information; and more collaborators coming together to learn, use, and create new content. There is increasing recognition by businesses, research institutions, policy makers, and funders that digital content, thoughtfully managed, not only supports a thriving cultural heritage sector, but also contributes more broadly to positive job creation and international competitive advantage. More has been the theme of digital stewardship as well. More work is being done to steward digital content than ever before. "Digital preservation makes headlines now, seemingly routinely. And the work performed by the community... is the bedrock underlying such high profile endeavors."[1] A recent example is the appearance of a since-deleted blog post in the Internet Archive Wayback Machine that could be evidence in the MH17 plane crash in Ukraine.[2] The current challenge of digital stewardship is managing, in a transparent and authentic way, this massively increasing volume of the digital content at levels of rapid upward scalability that require innovative approaches to management.

Last year the White House issued a major directive requiring agencies to increase open access to publications and data from federally funded research;[3] the National Institutes of Health, the world's largest public funder of research, launched a major new program focused on the use

[1] Kirschenbaum, M. (2014, July 22). *Software, It's a Thing*. Presentation at Digital Preservation 2014. Retrieved from https://medium.com/@mkirschenbaum/software-its-a-thing-a550448d0ed3

[2] Taylor, N. (July 28, 2014). The MH17 Crash and Selective Web Archiving. The Signal Blog. Library of Congress. Retrieved: from http://blogs.loc.gov/digitalpreservation/2014/07/21503/

[3] Holdren, J. P. (2013). "Increasing Access to the Results of Federally Funded Scientific Research." *Office of Science and Technology Policy*. Retrieved from http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

and management of big data, and appointed its first head of data science;[4] and the long-running case addressing Google's scanning of millions of books was decided, allowing scanning, searching, and other "fair uses" to continue unimpeded.[5] Commercial services such as Amazon Glacier, Rosetta, and Preservica and community-based platforms such as LOCKSS, DuraCloud, and the Digital Preservation Network continue to develop substantial functionality in support of medium and long-term stewardship. In the research area, EU funding of projects such as SCAPE and 4C are advancing our knowledge of effective digital preservation.

The first iteration of the *Agenda*[6] was released for 2014 and attracted significant attention from the digital preservation community. It identified opportunities and recommendations for addressing the most pressing technical, institutional, legal, and economic challenges faced by the digital preservation community.

The 2015 edition of the *Agenda* builds on the earlier work, updating the 2014 text, and identifying high-level action recommendations, directed at funders, researchers and organizational leaders, that will advance the community capacity for digital preservation, the evidence base for efficient and reliable practice, and the network of durable content that is available to the nation.

The 2015 *Agenda* is another step on a continuum towards successful stewardship. It includes specific actions that can be taken now, recognizing that it will require years of reflection, iteration, and refinement to identify comprehensive and effective interventions for the entire breadth of systems involved in stewardship—and it will require the contribution of the entire stewardship community to enact these interventions.

**Key Issues in Building Digital Content Collections**

Much of the investment and effort in the field of digital preservation has been focused on developing technical infrastructure, networks of partnerships, education and training, and establishing standards and practices. Little has been invested in understanding how the stewardship community will coordinate the acquisition and management of born-digital materials in a systematic and public way. A gap is starting to emerge between the types of materials that are being created and used in our society and the types of materials that make their way into libraries and archives. The stewardship community must recognize this gap, understand why it exists, and determine how it could be addressed at local, regional, and national levels.

## Overarching Challenges of Digital Content

Both born-digital and digitized content present fundamental new issues to stewards tasked with ensuring meaningful long-term access to content. The need for effective digital stewardship is urgent, because content and the context that makes it meaningful is changing rapidly. Moreover, effective digital stewardship requires collaboration and coordination, because organizations rely on information beyond their institutional boundaries. Digital curation of the

[4] NIH Office of the Director. (2013, December 9). NIH Names Dr. Philip E. Bourne First Associate Director for Data Science." *News and Events*. Retrieved from http://www.nih.gov/news/health/dec2013/od-09.htm

[5] Miller, C.C., Bosman, J. Siding With Google, Judge Says Book Search Does Not Infringe Copyright. (2013, November 14). *New York Times*. Retrieved from http://www.nytimes.com/2013/11/15/business/media/judge-sides-with-google-on-book-scanning-suit.html

[6] National Agenda for Digital Stewardship. (2014) National Digital Stewardship Alliance. Retrieved from http://www.digitalpreservation.gov/ndsa/nationalagenda/index.html

content each organization uses is too large for any single organization to do, and avoiding the risks of loss depends on actions taken across the community.

The stewardship community needs to develop a broad evidence base describing the *practice and content of stewardship,* including identifying the types and collections of content that are being used by its members, who is taking responsibility for these collections, what organizations have the capability to take on additional stewardship, what organizations can provide long-term access to preserved materials, and identifying what stewards are doing to reduce the risk of loss inherent to digital information. This evidence base could be used to identify where we are doing well, where the gaps are, where the single points of failure are, and where the opportunities are to coordinate to reduce risk to important collections.

**Approaches to Content Selection at Scale**

Both libraries and archives have established concepts of selection and appraisal that are meant to guide curators in making these often subjective decisions. It is difficult to evaluate how well libraries, archives, and museums are collecting and preserving the large amounts of digital data that their users, patrons, researchers, and institutions rely on. Traditional forms of scholarship like articles, edited volumes, and monographs (and their digital equivalents) are fairly well-understood in terms of how they fit into an institution's collection strategy. The non-traditional forms of evidence, like much of the data found on the open web, do not easily fit into existing acquisition processes. In addition, the usage data and logs that augment the open web data are becoming just as significant to researchers. The following recommendations and the discussion later in the document are meant to advance a fuller understanding of approaches to selecting in the digital environment.

**Core digital content recommendations:**

- *Build the evidence base for evaluating at-risk, large-scale digital content for acquisition.* Develop content environmental scans in each area of interest to the community that identify important collections, and the risks and efforts to ensure durable access to them.
- *Understand the technical implications of acquiring large-scale digital content.* Extend systematic surveys and environmental scans on preservation storage practices and organizational capacity to guide selection.
- *Share information about what content is being collected and what level of access is provided.* Communicate and coordinate collection priority statements at national, regional, and institutional levels.
- *Support partnerships, donations and agreements with creators and owners of digital content and stewards.* Connect with communities across commercial, nonprofit, private, and public sectors that create digital content to leverage their incentives to preserve.

## Organizational Policies and Practices

Despite continued preservation mandates and over ten years of work and progress in building a comprehensive practice around digital preservation, the community still struggles with advocating for resources, adequate staffing, and articulating the shared responsibility for stewardship. Underlying all of these challenges is a lack of prioritization of digital preservation programs. Integrating digital stewardship practice and thinking across an entire organization is a

core challenge, especially in a time of restricted resources. Part of the challenge is giving decision makers the information they need to make informed decisions and manage organizations that steward digital materials. We must also remember that a significant part of our work is related to and directly impacts commercial digital stewardship as well, so every effort that is created within the "for-profit" community should be shared among all constituencies when applicable. Efforts in the area of organizational roles and policies for digital stewardship should be focused on the following objectives. These are actions for which practitioners, managers, stakeholders, and funders can advocate and implement, as they work toward an environment where the mandate and need for digital preservation are matched with the resources, staffing, and an effective professional community prepared to meet those mandates and needs.

**Core Recommendations**

- *Advocate for resources.* Share strategies and develop unified messages to advocate for funding and resources; share cost information and models; and develop tools and strategies that inform the evaluation and management of digital collection value and usage.
- *Enhance staffing and training.* Explore and expand models of support that provide interdisciplinary and practical experiences for emerging professionals and apply those models to programs for established professionals. Evaluate and articulate both the broad mix of roles and the specialized set of skills in which digital stewardship professionals are involved.
- *Foster multi-institutional collaboration.* Foster collaboration through open source software development; information sharing on staffing and resources; coordination on content selection; and coordinated engagement with the development of standards and practices; and identify, understand and connect with stakeholders that may be outside of the cultural heritage sector.

## Technical Infrastructure Development

Broadly speaking, the infrastructure that enables digital preservation involves the staff, workflows, resources, equipment, and policies that ensure long term access to digital information. This section focuses specifically on the technical component of that infrastructure. Technical infrastructure can be generally defined as the set of interconnected technical elements that provide a framework for supporting an entire structure of design, development, deployment, and documentation in service of applications, systems, and tools for digital preservation. This includes hardware, software, and systems. Organizational policies, practices, and regulations inform many of the observations and recommendations for the development of digital stewardship technical infrastructure.

**Core recommendations:**

- *Coordinate and sustain an ecosystem of shared services*. Better identify and implement processes to maintain key software platforms, tools and services; to identify technologies which integrate well to form a sustainable digital workflow. And better models to support long-term sustainability for common goods are needed.
- *Foster best practice development.* Give priority to the development of standards, best

practices, especially in the areas of format migrations and long term data integrity

## Research Priorities

Research is critical to the advancement of both basic understanding and the effective practice of digital preservation. And research in digital preservation is under-resourced -- in part this is because the payoff from long-term access occurs primarily in the medium-long term and tends to benefit broad and diverse communities. The investment in the following two areas will yield unusually large impact:

**Core recommendations:**

- *Build the evidence base for digital preservation*. Give priority to programs that systematically contribute to the the overall cumulative evidence base for digital preservation practice and resulting outcomes -- including supporting testbeds for systematic comparison of preservation practices
- *Better integrate research and practice*. Give priority to programs that rigorously integrate research and practice or that increase the scalability of digital stewardship

A common challenge running through this report is the limited amount of empirical evidence available. The digital preservation community is beginning to develop a shared evidence base that can be used to answer these and similar questions; however, these studies must be broadened and repeated over time to establish a robust evidence base from which generalizable guidance can be drawn. Furthermore, decision makers should recognize that basic research in these areas often needs to be paired with the development, support, and evaluation of infrastructure.

### ABOUT THE NATIONAL DIGITAL STEWARDSHIP ALLIANCE

Founded in 2010, the National Digital Stewardship Alliance (NDSA) is a consortium of institutions that are committed to the long-term preservation of digital information. NDSA's mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. The NDSA comprises over 160 participating institutional members. These members come from 45 states and include universities, consortia, professional societies, commercial businesses, professional associations, and government agencies at the federal, state, and local level. NDSA organizations have proven themselves committed to long-term preservation of digital information.

To learn more about the NDSA: http://www.ndsa.org

### ABOUT THE AUTHORS

The joint leadership group of the NDSA authored the report and engaged in discussions to identify significant trends and challenges. The membership of the NDSA contributed to these discussions. This dialog was enriched by an extensive range of resources and current research.

The joint leadership group is made up of the Coordinating Committee members, the Working Group co-chairs, and the NDSA facilitator:

Micah Altman (MIT), Jefferson Bailey (Internet Archive), Karen Cariani (WGBH), Jim Corridan (Indiana Commission on Public Records), Jonathan Crabtree (UNC, Chapel Hill), Michelle Gallinger (Library of Congress), Andrea Goethals (Harvard Library), Abbie Grotke (Library of Congress), Cathy Hartman (University of North Texas), Butch Lazorchak (Library of Congress), Jane Mandelbaum (Library of Congress), Carol Minton Morris (DuraSpace), Kate Murray (Library of Congress), Trevor Owens (Library of Congress), Megan Phillips (NARA), Abigail Potter (Library of Congress), Robin Ruggaber (University of Virginia), John Spencer (BMS/Chace). Helen Tibbo (UNC Chapel Hill), Kate Wittenberg (Portico)

# 1. Introduction

Our culture is a digital culture. The photographs of our families, the communities where we share and receive news, the maps that give us new insight on where we're going and how to get there, the films and music that shape our shared experiences—now almost all digital. Our digital creations represent an investment in time, energy, and resources that require responsible care to remain viable over time. Effective digital preservation is vital to maintaining the authentic public records necessary for understanding and evaluating government actions; the verifiable scientific evidence base for reproducing research, and building on prior knowledge; and the integrity of the nation's cultural heritage. Substantial work is needed to ensure that today's valuable digital content remains accessible, useful, and comprehensible in the future — supporting a thriving economy, a robust democracy, and a rich cultural heritage.

The *National Agenda for Digital Stewardship* provides funders, decision-makers, and practitioners with insight into emerging technological trends, gaps in digital stewardship capacity, and key areas for research and development to support the work needed to ensure that today's valuable digital content remains accessible, useful, and comprehensible in the future to support a thriving economy, a robust democracy, and a rich cultural heritage.

The last ten years have seen strong global trends in the production and use of digital content. The theme of the decade has been *more*: more information being produced; more content being published and shared; more forms of publication and filtering; more public access to information; and more collaborators coming together to learn, use, and create new content. There is increasing recognition by businesses, research institutions, policy makers, and funders that thoughtfully managed digital content supports a thriving cultural heritage sector and contributes much more broadly to positive job creation and international competitive advantage. The challenge of digital stewardship is the challenge of managing this massively increasing volume of the digital content at scale. The rapid scalability needed requires innovative approaches to management of digital content.

Specific sectors mirror the general trend. In higher education there has been an explosion of new learners taking advantage of digital content through massive open online courses and other online educational systems, while at the same time, the production of digital content through social media has become ubiquitous, and data of all types is dramatically expanding in availability and importance.[7]

Governmental and legal actions also have a direct impact on digital stewardship. During the last year, the White House issued a major directive requiring agencies to increase open access to publications and data from federally funded research,[8] the National Institutes of Health, the world's largest public funder of research, launched a major new program focused on the use and

[7] Manyika, J. (2013). "Open Data: Unlocking Innovation and Performance with Liquid Information." *McKinsey Global Institute*.; Schwab, K., et al. "Personal data: The emergence of a new asset class." *An Initiative of the World Economic Forum*. 2011. http://www3.weforum.org/docs/WEF_ITTC_PersonalDataNewAsset_Report_2011.pdf

[8] Holdren, J. P. (2013). "Increasing Access to the Results of Federally Funded Scientific Research." *Office of Science and Technology Policy*. Retrieved from: http://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

management of big data and appointed its first head of data science,[9] and the long-running case addressing Google's scanning of millions of books was decided – allowing scanning, searching, and other "fair uses" to continue unimpeded.[10] Commercial services such as Amazon Glacier, Rosetta, and Preservica ; and community-based platforms such as LOCKSS, Duraspace, and the Digital Preservation Network continue to develop substantial functionality in support of medium and long-term stewardship. In the research area, EU funding of projects such as APARSEN and 4C are advancing our knowledge of effective digital preservation; and NIH's launch of its Big Data 2 Knowledge (BD2K) is beginning to spark intense interest in research in the stewardship of biomedical research.

The overall trend is clear: more information, in more forms, created by more people. This drives the need for digital stewardship practices to scale up accordingly.

The inaugural edition of the *Agenda* was released in 2014 and attracted significant attention from the digital preservation community. It analyzed the systems of digital stewardship and aimed to understand and conceptualize optimal (or at least much-improved) technical, institutional, legal, economic, and research systems.The 2015 edition of the Agenda builds on this work, updates the 2014 text, and identifies high-level action recommendations, directed at funders and organizational leaders, that will advance the community capacity for digital preservation, the evidence base for efficient and reliable practice, and the network of durable content that is available to the nation.

The 2015 *Agenda* is another step on a continuum towards successful stewardship. It includes specific actions that can be taken now, we recognize that it will require years of reflection, iteration, and refinement to identify comprehensive and effective interventions for the entire breadth of systems involved in stewardship--and it will require the contribution of the entire stewardship community to enact these interventions. The *Agenda* outlines the challenges and opportunities related to digital stewardship activities in four broad areas: Organizational Policies and Practices, Key Issues in Digital Collection Building, Technical Infrastructure Development, and Research Priorities. Each section articulates priority challenges, and then offers a set of *Actionable Recommendations* to address the challenges.

The *Agenda* is released under the auspices of the National Digital Stewardship Alliance, a membership organization of leading government, academic, nonprofit, and private sector organizations with digital stewardship responsibilities. Library of Congress provides essential organizational support, logistical support, and substantive collaboration and expertise. Members of the NDSA collaborate to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations.

### 1.1 NDSA Role

The NDSA continues to contribute to securing and broadening access to the expanding digital resources of the United States of America, develops and coordinates sustainable infrastructures for the preservation of digital content, advocates standards for the stewardship of digital objects, builds a community of practice, promotes innovation, facilitates cooperation

[9] NIH Office of the Director. (2013, December 9). NIH Names Dr. Philip E. Bourne First Associate Director for Data Science." *News and Events*. Retrieved from http://www.nih.gov/news/health/dec2013/od-09.htm

[10] Miller, C.C., Bosman, J. Siding With Google, Judge Says Book Search Does Not Infringe Copyright. (2013, November 14). *New York Times*. Retrieved from http://www.nytimes.com/2013/11/15/business/media/judge-sides-with-google-on-book-scanning-suit.html

between previously unaligned sectors, and raises awareness of the enduring value of digital resources and the need for active stewardship.

The Agenda described the challenges and opportunities for the entire stewardship community, of which NDSA is only a part. NDSA as a voluntary organization is primarily to provide guidance and coordination and, does not lay claim to enact the actionable recommendations in this document.

Notwithstanding, the priorities revealed in the Agenda provide a framework for prioritizing NDSA working groups, projects and collaborations. And NDSA is contributing in key areas to moving this community agenda forward – by building a broad and systematic evidence base on community preservation practices; capacity; and, in targeted areas, to survey community content at risk. Further, the NDSA membership collectively represents a community information reserve of over 100 petabytes of content that has been expertly selected and intensively curated for access to broad communities over the long term

With its national focus, the NDSA is in a unique position to identify and communicate the challenges, opportunities, and priorities for digital stewardship activity in the United States. The NDSA joint leadership group, digital stewardship experts elected from a cross-section of diverse sectors of the U.S. economy, including libraries and archives, academic, technology and commercial concerns, authored this strategic agenda.

## 2. Key Issues in Building Digital Content Collections

Much of the investment and effort in the field of digital preservation has been focused on developing technical infrastructure, networks of partnerships, education and training, and the establishment of standards and practices. Little has been invested in understanding how the stewardship community will coordinate the acquisition and management of born-digital materials in systemic way. A gap is starting to emerge between the types of materials that are being created and used in our society and the materials that make their way into libraries and archives. The community must recognize this gap, understand why it exists and how it could be addressed at local, regional, and national levels.

There are a large number of issues in building digital content collections. Many of the key issues are related to engaging with pivotal communities of practice, or to curating high-priority content types. However, there are overarching issues—the way in which the transition from physical to digital content has created fundamental new challenges, and the resulting urgent need for action that is coordinated across organizations.

### *2.1 Overarching Issues with Digital Content*

The need for effective digital stewardship is urgent because content and the context that makes it meaningful are changing rapidly. Moreover, effective digital stewardship requires collaboration and coordination, because organizations rely on information beyond their boundaries. The job of digital curation is too large for any one organization to do, and the risks of loss depends on actions taken across the community.

Coordination takes time, and there are many specific challenges to external and internal coordination, as we discuss in the *Organizational Practices section.* Notwithstanding, there are a

number of clear, vital, and actionable steps that can be taken towards improving coordination and its impact.

The stewardship community needs to develop a broad evidence base describing the *practice and content of stewardship,* including identifying the types and collections of content that are being used by its members, who is taking responsibility for these collection, what organizations have the capability to take on additional stewardship, and what stewards are doing to reduce risks. This evidence base could be used to identify where we are doing well, where the gaps are, and where there are opportunities to coordinate to reduce risk to important collections.

NDSA has made a number of concrete steps toward establishing this evidence base. The preservation storage survey[11] provides systematic information on what stewards are doing to reduce risk to content they curate, the staffing survey[12] provides systematic information on the resources available for curation, and content surveys, such as the web archiving survey[13], provide the beginnings of a map of stewarded content. Other projects such as the Keeper's Registry[14] , and the Memento[15] project are also growing this body of evidence. But much more is needed.

**Actionable Recommendations**

- Develop content scans in each area of interest to the community that identify important collections and the efforts to ensure durable access to them
- Continue to build systematic longitudinal evidence on the practice and content of preservation.
- Extend systematic surveys and scans on preservation storage practices and organizational capacity to guide selection

### *2.2 Approaches to Content Selection at Scale*

Collecting born-digital materials differs significantly from collecting analog materials. Both libraries and archives have established concepts of selection and appraisal that are meant to guide curators in making these often subjective decisions. Many of these concepts are still applicable, especially appraisal, which is useful when considering large volumes of material.[16]

---

[11] Altman, M., Bailey, J., Cariani, K., Gallinger, M., Mandelbaum, J., Owens, T. (May/June 2013) NDSA Storage Report: Reflections on National Digital Stewardship Alliance member Approaches to Preservation Storage Technologies. D-Lib Magazine. Vol. 19, Number ⅚. doi:10.1045/may2013-altman

[12] Atkins, W., Goethals, A., Kussmann, A., Phillips, M., Vardigan, (December 2013) M. Staffing for Effective Digital Preservation: An NDSA Report. National Digital Stewardship Alliance. Retrieved from: http://digitalpreservation.gov/ndsa/documents/NDSA-Staffing-Survey-Report-Final122013.pdf

[13] National Digital Stewardship Alliance Content Working Group. (June 19, 2012) Web Archiving Survey Report. National Digital Stewardship Alliance. Retrieved from: http://digitalpreservation.gov/ndsa/working_groups/documents/ndsa_web_archiving_survey_report_2012.pdf

[14] Burnhill, Peter (2013) Tales from The Keepers Registry: Serial Issues About Archiving & the Web. *Serials Review,* 39 (1), March 2013, pp. 3–20. http://dx.doi.org/10.1016/j.serrev.2013.02.003. Publisher's final copy of the work is also online at http://www.era.lib.ed.ac.uk/handle/1842/6682.

[15] Van de Sompel, H., Nelson, M. L., Sanderson, R., Balakireva, L. L., Ainsworth, S., Shankar, H. (November 2009). Memento: Time Travel for the Web. zrXiv preprint. arXiv:0911.1112.

[16] In an archival context, appraisal is the process of determining whether records and other materials have permanent (archival) value. Appraisal may be done at the collection, creator, series, file, or item level. Appraisal

And, as in the analog tradition, acquisitions should be based on local priorities, strengths, and documented policies. Still, it is difficult to evaluate how well libraries, archives, and museums are collecting and preserving the large amounts of digital data that their users, patrons, researchers, and institutions rely on. Traditional forms of scholarship like articles, edited volumes, and monographs (and their digital equivalents) are fairly well-understood in terms of how they fit into an institution's collection strategy. The non-traditional forms of evidence, like much of the data found on the open web, do not easily fit into existing acquisition processes. In addition, the usage data and logs that augment the open web data is becoming just as significant to researchers.[17]

The character of digital data also complicates its collection and hence its preservation. Digital data is incredibly easy to create, replicate, and share. The ownership and provenance of data created on the web is often unknown or unclear. The size and dynamism of data is ever increasing, and the granularity and interconnectedness ever more complex. Although much is made of the digital trace that can be left online,[18] research indicates that a significant amount of the data that permeates nearly all aspects of life, culture, and scholarship today will not be available at a library or archives[19] unless attention and priority is paid to actively collecting digital materials.

Selection has always been about making collection decisions that align with strengths and missions of an institution. The British Library recently released a revised collection strategy[20] that reflects a focus on born-digital materials. Special and rare collections are often the marquee collections in an analog environment, the uniqueness and value of special collections sets one library apart from another. The strategy to actively acquire unique born-digital materials, like web archives or digital records, documents and hard drives for manuscript archives, continues the strengthening of special collections. Stewardship organization should not shy away from collecting born-digital materials that lack a predetermined process for acquisition, especially when data may be at-risk or without a custodial home. Also required is better integration and better knowledge about how digital collecting supports the mission of stewardship organizations, where gaps and opportunities in the national landscape of digital collections exist and how the long-term preservation of digital collections impact technical architectures.

### *2.2.1 Connection to Researchers*

Related to the challenge of selection is how digital collections interact with users. Often, researchers increasingly desire not only access but enhanced use options and tools for engaging with digital content. Usability is increasingly a fundamental driver of support for preservation,

---

can take place prior to donation and prior to physical transfer, at or after accessioning. Society of American Archivists. *Glossary of Archival and Records Terminology*. July 14, 2014. http://www2.archivists.org/glossary/terms/a/appraisal

[17] Liao, H. and Petzold, T., (August 27-29, 2014) Geographic and linguistic normalization: towards a better understanding of the geolinguistic dynamics of knowledge. I . OpenSym '14. A.. ACM 978-1-4503-301 6-9/14/08. http://sx.doi.om/10/1145/2641580.2641623.

[18] Rosen, Jeffrey. The Right to Be Forgotten. *Stanford Law Review Online. 64; 88. February 13, 2012.* http://www.stanfordlawreview.org/online/privacy-paradox/right-to-be-forgotten?em_x=22

[19] McCown, Frank; Sheffan Chan, Michael L. Nelson, Johan Bollen. 5th International Web Archiving Workshop 2005 Proceedings. http://iwaw.europarchive.org/05/papers/iwaw05-mccown1.pdf

[20] From Stored Knowledge to Smart Knowledge: The British Library's Content Strategy 2013-2015. http://www.bl.uk/aboutus/stratpolprog/contstrat/british_library_content_strategy_2013.pdf

particularly for ongoing monetary support. Models for access continue to evolve as methods for analyzing and studying contemporary born-digital and historic digitized materials are available.

A number of experiments around the research use of web archives are instructive to stewards of any kind of digital content. Researchers in Europe[21] are exploring the methods to create a corpus of web sites and web archives via an automated process that makes it possible to track versions, annotate and analyze the data while keeping it in a stable state so it is possible to compare results over time. Web researchers and curators have also experimented with tools[22] that allow researcher to select and archive sites from the live web to build a corpus for ongoing analysis.

The practice of providing access to data sets at the data analysis level is still rare but proposals are coming to the surface that take advantage of existing infrastructure[23] to serve researchers the raw data they want. More can be done to establish new practices and share knowledge around the gap between user and researcher needs for digital content and the access models that are currently available.

*2.2.2 Connection to Creator Community*

Digital stewardship organizations need to leverage activities in content-creating communities that generate incentives to preserve that may be unrelated to the specific interests of the stewardship community but generate positive stewardship benefits.[24]

For example, the Uniform Electronic Legal Material Act[25] establishes an outcomes-based, technology-neutral framework for providing online legal material with the same level of trustworthiness traditionally provided by publication in a law book. It requires official electronic legal material to be authenticated, preserved, and made accessible, providing strong incentives for state governments to allocate resources to effectively steward their digital data. UELMA efforts are driven by legislative dynamics but have been influenced by library and archive practice. UELMA has the potential to positively impact digital stewardship practice, and stewarding organizations in the states should be familiar with UELMA and advocate for its passage.

Stewardship has also made significant inroads in the creative content communities. The Motion Picture industry has published two "Digital Dilemma" reports[26] on the challenges facing

---

[21] Brügger, N. (April 25, 2013). Fundamental tools for web archive research. Scholarly Use of Web Archives Open IIPC Conference. Ljubljana, Slovenia. Retrieved from: http://netpreserve.org/sites/default/files/resources/Scholarly%20Use%20of%20Web%20Archives%20kopi.pdf

[22] LiveArchiving Proxy. INA-DLWeb. Retrieved August, 26, 2014 from https://github.com/INA-DLWeb/LiveArchivingProxy

[23] Markman, C. & Zavras, C. (March/April 2014). BitTorrent and Libraries: Cooperative Data Publishing, Management and Discovery. D-Lib Magazine. Vol. 20, No. 3-4. http://www.dlib.org/dlib/march14/markman/03markman.html

[24] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (February 2010). Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Retrieved from http://brtf.sdsc.edu/

[25] Electronic Legal Material Act. (2011). Uniform Law Commission. The National Conference of Commissioners on Uniform State Laws. Retrieved from: http://www.uniformlaws.org/Act.aspx?title=Electronic%20Legal%20Material%20Act

[26] The Science and Technology Council of the Academy of Motion Picture Arts and Sciences. (2007). The Digital Dilemma: Strategic Issues in Archiving and Accessing Digital Motion Picture Materials. http://www.oscars.org/science-technology/council/projects/digitaldilemma/

major studios and independent filmmakers in preserving their digital audiovisual materials. These reports suggest next steps for engagement with these communities to implement best practices for digital multimedia content stewardship.

In the digital audio industry, solutions supported by the digital stewardship community are now making their way into industry practice. The "Metadata Schema Development for Recorded Sound" project[27] focused on creating a standardized approach for gathering and managing metadata for recorded music and developed a software tool (the Content Creator Data Tool[28] to assist creators and owners in collecting the data. These efforts are being widely adopted in industry organizations such as the Music Business Association's Digital Asset Management Workgroup[29] and the Recording Academy, Producers and Engineers Wing.

While the geoarchiving community remains small, it has developed a body of research[30] that can direct future efforts, should support for further investigation be made available. This includes efforts to leverage the Federal government Geospatial Platform[31] activities and the Federal Geographic Data Committee Circular A-16 Supplemental Guidance[32] implementation efforts.

Each of these industries has identified incentives that drive the stewardship of their digital materials. These incentives may align with the interests of collecting organizations, but in many cases they operate independently of them. These "industry" incentives may be much stronger than those of the stewarding community because of the competitive business advantage they supply the participants. In some instances, proper stewardship may determine whether an industry continues to thrive or even survive. The stewardship community should explore every possible opportunity to leverage these strong incentives.

The December 2012 release of the Library of Congress "National Recording Preservation Plan"[33] offers an example. In addition to proposing industry-wide recommendations on building the national sound recording preservation infrastructure, it suggests blueprints for implementing preservation strategies (both analog and digital) and promoting preservation efforts in the service of educational purposes. This approach could be elaborated upon for other types of digital content, with organizations such as the NDSA coordinating national approaches to preservation strategies for multiple content types, as long as the creating industries are deeply engaged in the process.

**Actionable Recommendations**

- Support the ongoing evaluation of digital collections and their impacts.
- Communicate and coordinate collection priority statements at national, regional, and institutional levels.
- Explore privacy issues in born-digital collecting.

---

[27] Metadata Schema Development for Recorded Sound. Library of Congress. Retrieved August 26, 2014 from http://digitalpreservation.gov/partners/bms_chace.html

[28] Content Creator Data project. Retrieved August 26, 2014 from http://ccddata.com/

[29] http://musicbiz.org/sectors/information-technology-sector

[30] Link to NDSA Geospatial quick reference when available

[31] https://www.geoplatform.gov/

[32] https://www.fgdc.gov/policyandplanning/a-16

[33] http://www.loc.gov/rr/record/nrpb/PLAN%20pdf.pdf

- Develop further understanding and proficiency in the tools researchers want to interact with digital collections.
- Connect with the communities across commercial, nonprofit, private, and public sectors that create digital content to leverage their incentives to preserve.

### 2.3 Content-Specific Challenges

In addition to the cross-cutting issues discussed above, specific forms of content pose urgent challenges to stewardship. Scientific data sets, dynamic web content, software, and massive collections of recorded video and audio pose specific technical, institutional questions that go beyond issues of the scale of content. This content is increasingly being recognized as a vital part of the scientific, cultural, and public record—but remains at high risk of loss.

*2.3.1 Organizing and Ensuring Long Term Access to Scientific Data Sets*

Some of the most acute challenges of digital content can be illustrated by considering the curation of digital research data. The sheer *scale* of research data represents a daunting curatorial task. With newly developed scientific instrumentation and the growing use of computer simulations, a research team can generate many terabytes of data per day. Data curators face management at the petabyte scale (1,000 terabytes) and well beyond. Scientific fields such as particle physics, with its collider data, and astronomy, with its sky surveys, as well as research fields and methods like bioinformatics, crystallography, and engineering design generate massive amounts of digital data.

Although some research data are no more complex than other objects that are routinely curated, a portion of digital research data are complicated to curate. Research data can be heterogeneous, ranging from numeric and image based, to textual, geospatial, and other forms. There are many different information standards used (and not used), as well as many different approaches to information structure (e.g., XML-structured documents vs. fixed image and textual file formats). Moreover, the research communities that produce data are equally diverse; data management practices vary greatly both within and between disciplines. There may also be commercial interests in the data and associated data practices.

Perhaps the overriding challenges in all respects to digital research data are the affiliated costs. Domain researchers, technologists, information scientists, and policymakers are searching for sustainable economic models with the ability to accurately predict costs and to balance them across the lifecycle (e.g., costs for ingest, archival management, and dissemination), and through federated inter-institutional repository systems. Managing research data will also require stewards to take on new roles. These may include enabling researchers to curate their research, absent professional expertise, applying and adapting metadata in new ways, and collaborating with researchers in developing new workflows, models, and tools.

One of the biggest needs for research data is a records schedule that reflects an understanding of the variations in data, be them raw, processed, summary, aggregate, preliminary, public use or metadata.[34] Another is a clearer understanding of the maze of data uses, reuses, incentives, mandates and responsibilities. There is no one-size-fits-all approach

[34] There are commonalities across data based on the methodology that was used to create the data (and differences by domain, but exploring commonalities based on methodology could produce important results) – DCC's DAF codified archival appraisal without referring to it either – this is significant because there is a lot of archival appraisal and records management literature that could be helpful

when it comes to resolving the management challenges of research data; however,[35] progress might be made by mobilizing the digital preservation and curation community toward in-depth study of these challenges of scale, complexity, research community practice, and cost with the aim of developing new recommendations and potential long-term solutions.

Aside from needs for evidence-based research, projects like California Digital Library's DataUp tool, which work to make it easier for researchers to make their data legible, reusable, and easy to submit into repository systems, are clearly needed to inform practice.[36] In this respect, there is a need for a range of related efforts that work to help bridge the gap between the practices of working researchers' data management into long term stewardship.

**Actionable Recommendations**

- Support more research-based practice and practice-based research for scientific data preservation.
- Support tool development in science data preservation.

*2.3.2 Dynamic and Heterogeneous Web and Social Media*

The 2014 National Agenda cites Web and Social Media as an area of concern for preservation. Because of the growth in quantity and complexity of this content type, it is again included the 2015 agenda with a deeper description of the problems faced in preserving this content. The stewardship community must make connections with other communities to tackle the web preservation problem from the other side, by making the case to web content producers that archivability is a criterion worth considering alongside accessibility, performance, SEO, standards compliance, and usability.

The Web has changed considerably in the 25 years since its inception, to say nothing of the 18 years since cultural heritage organizations started to more systematically preserve it. The scale of the Web has grown exponentially, and what was once a network of discrete hypertext documents has given way to an "executable" environment characterized by interactive web services. The mainstay tools of the web preservation community—the Heritrix archival crawler and the Wayback replay platform—are unfortunately maladapted to the contemporary Web. The challenge of scale confounds solutions that compensate for some of these tools' shortcomings as well as underscores the fractional amount of web content that even a dedicated and growing community can effectively preserve.

Member institutions of the International Internet Preservation Consortium (IIPC) developed Heritrix and Wayback and are presently working to develop a community to prospectively stabilize, enhance, and support these fundamental open source tools. Involvement by the broader web preservation community is not just needed on extending the capabilities of these and other web archiving technologies but also in generalizing their accessibility and ease-of-use, elevating consideration of archivability by web content producers, connecting web

[35] Lyon, Liz. "Dealing with Data: Roles, Rights, Responsibilities and Relationships. Consultancy Report." (2007); Graham Prior, 2012 Managing Research Data, Facet Press.

[36] The DMPTool and DataUp:Helping Researchers Manage, Archive, and Share their Data. Strasser & Cruse (2013) http://rdmi.uchicago.edu/sites/rdmi.uchicago.edu/files/uploads/Strasser,%20C%20and%20Cruse,%20P_The%20DMPTool%20and%20DataUP-Helping%20Researchers%20Manage,%20Archive,%20and%20Share%20their%20Data.pdf

archive selection and access to tangible use cases, and developing policies and best practices that keep pace with the changing Web. The resource-intensive of web preservation is a challenge for continuing to train web archiving specialists and to grow the community of practice.

There is growing interest among web archive end-users to be able to explore and interact with web archives in ways beyond browsing them through Wayback. Commercial search engines have engendered expectations for full-text search of web archives, whose size and temporal aspect present a major engineering challenge. Researchers want to apply computational approaches to raw web archive data. Such demands require increased collaboration to understand and cultivate use cases, new analytical tools and methods, new IT infrastructures, and policy experimentation. The expanding Memento[37] service which can reveal content from several web archives while a user is browsing the live web will help surface the extent of distributed web archives and more seamlessly integrate historical content into the contemporary Web.

Web preservation efforts are increasingly concerned with social media content. Social media epitomizes the mismatch between the contemporary Web and the web crawling paradigm, exacerbates concerns over privacy and copyright, and highlights not-easily-resolvable tensions between corporate business models and cultural heritage organizations' interest in preserving and providing access to data. Application programming interfaces (APIs) represent a service provider-sanctioned method that overcomes some of these policy and technical challenges but also reflect a troubling shift from the adequacy of a general-purpose tool to the prospect of having to devise individualized strategies for each platform.

Social media companies are but some of many parties with whom coordination will be increasingly necessary to advance the cause of web preservation. Social science internet researchers also offer significant experience wrangling with the ethical and policy issues associated with social media data, and the digital stewardship community could learn from such efforts.

The leading edge of web archiving is rapidly converging with software preservation as the model of standalone, client-based software gives way to web-based platforms. This burgeoning class of software is built on familiar web technologies and has the general characteristics of databases, including some clear guidelines for preservation[38]. However, these tools are executed in the browser, updated frequently and opaquely and thus offer fewer affordances for third-party collection. The collection of content from these web sites will require deeper coordination with service providers than the cultural heritage community (or service providers) have been accustomed to.

By making connections with other communities, particularly content producers and researchers, and continuing to explore alternative capture and access methods, cultural heritage institutions will be better positioned to ensure that the results of our web preservation efforts meet the needs researchers in the future. In summary, while the investments in web harvesting and collecting tools have paid significant dividends already, as web content moves further and further from a document paradigm to a dynamic application paradigm we now find ourselves

[37] Van de Sompel, H., Nelson, M. L., Sanderson, R., Balakireva, L. L., Ainsworth, S., Shankar, H. (November 2009). Memento: Time Travel for the Web.arXiv:0911.1112.

[38] Ribeiro, C., Gabriel, D. (11 March 2009) Database Preservation Briefing Paper. Digital Preservation Europe. Retrieved from http://www.digitalpreservationeurope.eu/publications/briefs/database_preservation_ribiero_david.pdf

needing to retool and develop new approaches for dealing with much of the most popular web content.

**Actionable Recommendations**

- Invest in the development of web archiving tools that can capture dynamic web content and social media.
- Engage with the communities of web producers, service providers and creators, especially social media companies, to instill the value of archivability in their terms of use and design.
- Increase collaboration with researchers who use web archives to understand and cultivate use cases, new analytical tools and methods, new IT infrastructures, and policy experimentation.

### *2.3.3 Getting Serious about Software Preservation*

Software is simultaneously a baseline infrastructure and a mode of creative expression. It is both the key to accessing and making sense of digital objects and an increasingly important historical artifact in its own right. When historians write the social, political, economic, and cultural history of the 21st century they will need to consult the software of the times. As such, it is essential that the digital stewardship continue to make strides to ensure long term access to software. To date, there are a series of significant projects and programs focused on software preservation, however, this has largely been an activity of a few individual organizations and needs to be a key priority of a whole host of stewardship organizations.

Much of the groundwork for preservation in this area was laid in the Preserving Virtual Worlds NDIIPP-funded initiative. With that noted, this has moved from an area of research interest into a place in which considerable, progress has been made but which is ripe for considerable collection and infrastructure development. The work of the National Software Reference Library and its partnership with Stanford University to preserve The Stephen M. Cabrinety Collection in the History of Microcomputing illustrates how partnerships can help to get disk image copies of historical software. With that noted, as Matthew Kirschenbaum suggests in "An Executable Past: The Case for a National Software Registry" there are also critical reasons for the collection of source code as well. Beyond this, connected to the development of emulation platforms (like JSMESS[39] and Olive Library[40]), we are rapidly approaching a world in which it will be possible to make historical software collections replay-able over the web.

This progress is exhilarating; however, given the significance of software to our society, this work needs to be significantly scaled up. The National Digital Stewardship Alliance would like to see a range of institutions take the following actions.

**Actionable Recommendations**

[39] The JavaScriptMESS Project is a porting of the MESS Emulator. A program that emulates hundreds of machine types into the Javascript language. Retrieved August 20, 2014 from http://jsmess.textfiles.com/.

[40] Olive Executable Archive is a collaborative project seeking to establish a robust ecosystem for long-term preservation of software, games, and other executable content. Retrieved August 20, 2014 from https://olivearchive.org

- Organizations with a stake in long term access to software need to identify what stake they have in software preservation and begin to carve out and declare what kinds of software they intend to collect to the broader community.
- There is a need for outreach and engagement with the software industry in these problems and issues.
- Investments in research and tool development for virtualization and emulation of computing environments, like the Olive Executable Archive project and the JavaScript Multi Emulator Super System, are necessary to make software usable.
- Basic research is still required to inform the development of tools and infrastructure to support the preservation of entire computational environments and software that runs as web applications.

*2.3.4 Scale and Complexity of Moving Image and Recorded Sound Data*

Digital preservation and stewardship of motion picture film, audio, and video presents a multitude of challenges. There is a need for both new standards and for the evolution of existing standards, such as preservation-quality reformatting and a myriad of issues that arise from creating and managing large files—not only storage, but the long-term ability to manage and playback these files. While movie and recording industries *should* collaborate with cultural heritage institutions, this is not always the case. It is vital that both content creators and stewards work together to develop standards and workflows that will ensure long-term access to our recorded and moving image heritage. The cultural heritage community must continue to engage and encourage private/commercial and institutional relationships. In many cases, the tools and applications needed for both environments are useful for each community.

The ease of digital media creation has erupted from the multitude of easy-to-use cameras, each creating a file format output different from each other. Although ideally standards would be set that commercial vendors and those creating media equipment could adhere to, that is not the current situation. The digital preservation systems and infrastructure must be able to accommodate the ever-growing list of file formats to allow efficient preservation, access, and migration—whether that be transcoding to a single format or the ability to store and retrieve many native formats. That said, decisions and choices will need to be made as to what content is critical to keep. Although raw camera footage or the studio audio takes are in many, many formats, the file sizes are less daunting than final products like a feature film, or a TV program, or a master sound recording. These final products are very large media files, also in varying formats, that need to be moved around, stored, and preserved. Large files take more time to move, copy, process, and store. This means more resources for computing power, storage, people, and time.

Finally, the analog media created over the last 50-60 years is deteriorating at a rapid rate. Video tape and sound recording formats are becoming obsolete—the equipment needed to playback the formats are disappearing, and the physical tape itself is deteriorating. Many of the tape-based format playback machines are not available, have not been available for years, and require parts that are no longer manufactured. Digitizing analog materials is now considered the best preservation strategy and the best method for new distribution and access. This adds to the increasing collections of digital media files that require long-term preservation. This is a major catastrophe in process, and cultural heritage organizations must address it.

Rights issues are another complicating factor for preserving recorded sound and moving images. They are not covered currently under the exceptions in Section 108 of the Copyright Law, meaning all non-text-based works cannot explicitly be copied for preservation purposes, though they can be copied for fair use purposes[41]. For sound recordings alone, rights considerations need to be given to the recording itself, the performer, and the writer of the music. Films and televisions programs have distributors, copyright owners, talent unions, and owners of third party materials used in a final program or film—music, stills, historic footage. Where is the responsibility to preserve this content and provide it to the necessary agencies/unions if needed? Who can legally preserve it? Many funders will not help support preservation efforts unless public access is promised. How can an institution fulfill that promise when the rights issues for access are so complicated? Where are the gaps in this process that Congress should address, considering the divergent (and declining) revenue streams of content creators and content stewards and the fact Intellectual Property owners have strong lobbies?

Much of the 20th- and 21st-century cultural heritage and history is documented on audio visual media. As a democratic nation that sees the importance of understanding the past as we look to the future, it is important to find solutions for the long-term preservation, storage, and access of these materials. As a result, there is a significant need to further define and communicate what preservation formats are in this area and what, exactly, workflows should look like for working with and maintaining the authenticity of increasingly complex forms of digital audio and video files.

**Actionable Recommendations**

- Engage and encourage relationships between private and commercial and heritage organizations to work together to develop standards and workflows that will ensure long-term access to our recorded and moving image heritage.
- Support the ability of digital preservation systems and infrastructure to accommodate the ever-growing list of file formats to allow efficient preservation, access, and migration.
- Explore options for dealing with difficult Intellectual Property rights for preserving recorded and moving image materials.

### *2.3.5 Computational Techniques for Managing Records*

The potential loss of electronic records of business, organizations, and government, and the loss of the underlying information these records contain, poses a significant threat to the American memory.[42] Whether it's an electronic diary, an email exchange, or the documentation of government transactions, each of these records is at risk of disappearing unless thoughtful action is taken to preserve important information. Preserving electronic records efficiently and cost-effectively remains a tremendous challenge that needs to be addressed on many levels[43] The volume of records generated and held by individuals and institutions in electronic format

---

[41] Section 108 Study Group. (2008) Section 108 Study Group Report. Library of Congress. p 106. Retrieved from http://section108.gov/docs/Sec108StudyGroupReport.pdf

[42] *Future Watch: Strategies for Long-Term Preservation of Electronic Records.* Hoke, Gordon E J, CRM. Information Management Journal 46, 3 (May/Jun 2012): 26-28,30-31,47. Retreived through ProQuest http://search.proquest.com/docview/1019286317

[43] One attempt at addressing the problem is the Presidential Directive on Records Management http://www.whitehouse.gov/sites/default/files/omb/memoranda/2012/m-12-18.pdf

requires changes to traditional paper-based procedures. Rather than relying on file clerks to organize and store information, the information creator—each institution and individual—will be responsible for properly managing his or her own electronic records[44] A proper infrastructure, supplemented with public outreach, will be critical in educating the public about current deficiencies in long-term electronic preservation and in equipping them to properly save important materials.

The scale and heterogeneity of electronic records prompts a particular set of challenges. Where audio-visual content is increasingly pushing the limits of infrastructure in one kind of scale, with massive files, electronic records push issues of scale in a different way. In this case, it's an issue of massive numbers of relatively small files. In particular, there is a need to identify and scrub personally identifiable information and describe massive amounts of these files. As an example of this issue, 2012 report, Transforming Classification of the Public Interest Declassification Board described the rate of declassification using current processes with this example: "It is estimated that one intelligence agency would, therefore, require two million employees to review manually its one petabyte of information each year. Similarly, other agencies would hypothetically require millions more employees just to conduct their reviews[45]." To this end, there is a critical need for institutions to apply things like eDiscovery tools, natural language processing techniques and machine learning in the development of workflows and practices to enable the professional practices of archives and records management scale to work for electronic records.

**Actionable Recommendations**

- Support the application of automated workflows, eDiscovery tools, natural language processing techniques, and machine learning in the development of workflows and practices in dealing with voluminous and heterogeneous records.

# 3 Organizational Policies and Practices

Despite continued preservation mandates and over ten years of work and progress in building a professional practice around digital preservation, the community still struggles with advocating for resources, adequately staffing and articulating the shared responsibility for stewardship. Underlying all of these challenges is a lack of prioritization of digital preservation programs in institutions. The section below outlines some of the most pressing challenges and possible actions towards solutions that would raise the profile, prioritization, and effective management of digital stewardship actions.

## *3.1 Advocate for Resources*

There will never be enough resources to "save everything" stewarding organizations wish to preserve; this has always been the case. That said, cultural heritage organizations have new

[44] "to have an effective records management program, agency records management staff must have a baseline of knowledge about electronic records and how to manage them. Records staff do not need to be technological experts, but they have to understand certain fundamental principles and practices of managing electronic records." http://www.archives.gov/records-mgmt/resources/self-assessment-2011.pdf ; *2011Records Management Self-Assessment Report*, NARA.

[45] http://www.archives.gov/declassification/pidb/recommendations/transforming-classification.html

responsibilities to steward their growing digital collections on top of responsibilities to preserve and provide access to their analog collections. The economic downturn that started in late 2007 continues, and it will be years before many organizations return to earlier levels of funding. Still, stewarding organizations need to advocate for appropriate resources, and appropriate reallocation of resources, to tackle the task of digital stewardship. As such, stewarding organizations need to be able to offer value in exchange for the resources required to successfully address long-term digital stewardship issue.

Managers and stakeholders making resource decisions do so in an environment where estimating costs for digital stewardship is complex and not well understood. Transitioning digital stewardship costs from being supported by grants to being supported by regular budgets often begins with hiring staff (or adding duties to existing staff) directly responsible for digital preservation activities. To account for costs beyond staffing, numerous costing models exist but there are little comparative or longitudinal data to back up cost estimates. The European Union has funded a project called Collaboration to Clarify the Costs of Curation (4C)[46] that aims to help organizations invest more effectively in digital curation by providing a cost modeling tool and framework that explores the aspects of 'benefit,' 'risk,' 'value,' 'quality,' and 'sustainability.' It is analyzing previous work on cost modeling for digital preservation and building on the recommendations of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access.[47]

Preserving Digital Objects with Restricted Resources (POWRR)[48] is another project of note. It is aiming to understand the barriers and opportunities small- and medium-sized institutions face when they have been given digital preservation responsibilities but few resources to support digital preservation activities.[49] The project is evaluating digital preservation tools and services small- and medium-sized institutions could implement. A workshop series is being planned to disseminate findings and guidance from the project. Results from these projects will help to clarify costs and improve decision-making and strategic planning, which can in turn help to advance knowledge about the resources needed for the long-term management and development of sustainable infrastructure for digital preservation.

In addition to cost information and models, other methods of measuring and providing evidence of the value of digital stewardship activities is needed. Improved and sharable metrics about the quality and success of digital stewardship activities can help guide decision-making. Performance statistics have long been collected and compared to help libraries evaluate and improve the management of their collections and operations.[50] An effort is underway[51] to

---

[46] Kejser, U.B., Johansen, K.H.E., Thirifays, A. et. al. (2014, June 30). D3.1-Evaluation of Cost Models and Needs & Gaps Analysis. *Collaboration to Clarify the Cost of Curation.* Retrieved from http://www.4cproject.eu/

[47] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. (2010, February). Sustainable Economics for a Digital Planet: Ensuring Long-Term Access to Digital Information. Retrieved from http://brtf.sdsc.edu/

[48] Rinehart, A. K., & Prud'homme, P., & Huot, A. R. (2013). Overwhelmed to action: digital preservation challenges at the under-resourced institution. *OCLC Systems & Services*. 30(1). doi: 10.1108/OCLC-06-2013-0019. Retrieved from http://digitalpowrr.niu.edu/

[49] Schumacher, J. Digital POWRR Interim Report to Institute of Museum and Library Services. (2013). Retrieved from http://powrr-wiki.lib.niu.edu/images/b/bf/2013_Dec_Digital_POWRR_Interim_Report2.pdf

[50] Annual Library Statistics. (2014). *ARL Statistics*. Retrieved from http://www.arl.org/focus-areas/statistics-assessment

[51] Wacha, M., & Wisner, M. (2011). Measuring Value in Open Access Repositories. *The Serials Librarian*, 61(3-4), 377–388. doi:10.1080/0361526X.2011.580423. Retrieved from http://www.diglib.org/forums/2013forum/schedule/21-2/

determine ways of measuring collection usage across digital library platforms that inform management decision-making. This work should be broadened and built upon to give those responsible for digital stewardship the tools they need to advocate for the resources required.

Digital stewardship is important beyond the cultural heritage sector. Critical data are stewarded for scientific research authentication and repurpose; large data sets are mined for competitive advantage; data availability drives innovation. Stewardship in these areas is often driven by mandated actions, strategic tactics, or business models. These motivations are important to identify, understand, and apply, where appropriate, to the cultural heritage sector.

**Actionable Recommendations**

- Share the community learning from ongoing projects that help clarify costs, improve decision-making, and improve strategic planning for digital stewardship.
- Develop tools and strategies that inform the evaluation and management of digital collection value and usage.
- Identify and connect with stakeholders outside the cultural heritage sector to understand their motivations for digital stewardship and how that can inform NDSA efforts.

*3.2 Staffing and Training For Digital Stewardship*

Digital preservation professionals are often the intermediary between the information technology and curation communities. As the stewardship of digital materials becomes a responsibility for an increasing number and variety of institutions, education, training, and workforce development are key elements in supporting the expertise necessary for building a competent base of current and future of digital stewards. Key issues in this area include: exploring more practical, immersive internships, and fellowship for new professionals; the need for greater fluency with technologies across the field; more robust and accessible professional development opportunities; better understanding of career paths and organizational roles for digital curators and preservationists; affiliations with data management and preservation in non-humanities disciplines; and the exploration of collaborative opportunities between educational programs, students, and employers in the digital preservation community.

There have been significant efforts in digital preservation training and education that provide a baseline of information,[52] a core curriculum for professional development,[53] widely available workshops and training,[54] and opportunities for professional networking and knowledge sharing.[55] In late 2013 the NDSA Standards and Practices Working Group released a report of their analysis of their 2012 staffing survey.[56] In it, most respondents shared they expect their digital holdings to increase substantially with 20% expecting their digital holdings to double. They also indicated a need to nearly double the number of full-time equivalents (FTEs)

[52] Beagrie, N. & Jones, M. (2008, November). Preservation Management of Digital Materials: The Handbook. *Digital Preservation Coalition.* Retrieved from http://www.dpconline.org/advice/preservationhandbook

[53] Digital Preservation Management: Implementing Short-term Strategies for Long-term Problems. (2014). *MIT Libraries*. Retrieved from http://www.dpworkshop.org/workshops/fiveday.html

[54] Digital Preservation Outreach & Education. (2014). *Library of Congress.* Retrieved from http://digitalpreservation.gov/education/courses/index.html

[55] International Conference on Digital Preservation. (2014). Retrieved from http://ipres-conference.org/

[56] Arms, C., Chalfant, D., DeVorsey, K., Dietrich, C., Fleischhauer, C., Lazorchak, B. Morrissey, S., Murray, K. (2014, February). The Benefits and Risks of the PDF/A-3 File Format for Archival Institutions. *National Digital Stewardship Alliance.* Retrieved from http://hdl.loc.gov/loc.gdc/lcpub.2013655115.1

responsible for digital stewardship. To help meet this demand 75% of respondents report that existing staff are being retrained. In a list of the most preferred attributes of a digital preservation professional a "passion and motivation for digital preservation" and "knowledge of digital preservation standards, best practices, and tools" came out at the top while a "certificate or degree in digital preservation" and a "degree in computer science" were at the bottom. Genuine interest and motivation to learn about a subject cannot be taught in a workshop or training session; similarly, knowledge about standards and practices in an evolving field is best gained through direct, practical experience.

A recent successful model for providing practical experience in a dynamic environment is the IMLS-supported National Digital Stewardship Residency program.[57] It is a highly competitive residency for those who have recently finished a relevant master's degree to be placed in an academic, federal, non-profit, or cultural heritage organization to work on a digital stewardship program. The NDSR program expanded beyond the initial effort in the Washington D.C. region to cohorts based in Boston[58] and New York.[59] The model is worth noting: the resident hosts also went through a vetting process to select the most appropriate projects, the residents received two weeks of training prior to starting their project, they received support to attend conferences, and the cohort meet regularly throughout the residency. Many benefited from this model: all of the initial residents found full-time employment in their desired fields or went on to more advanced degrees; some of the host institutions were able to convert their participation in the program into justifications for new full-time positions; and the broader stewardship community gained experienced and engaged professionals who will become leaders in their fields. These types of opportunities should be shared with current professionals who want to gain skills and experience in digital preservation. In addition, longitudinal analysis of the effectiveness of the NDSR program and other types of education and training programs will be key to determining the best approach to meeting digital preservation staffing needs.

Related to education and training is the definition or classification of digital preservation positions. The NDSA Staffing Survey analysis reveals a wide range of activities, from research to cataloging to selection for preservation, that are included in digital preservation roles and that these activities are done in a variety of departments. These types of assessments of needed skills[60] may surface structural challenges to digital stewardship staffing. Digital stewardship requires a much broader individual skills mix than what has traditionally been needed across the range of cultural heritage institutions, but efforts to identify and advance the needed skills are still in their

[57] National Digital Stewardship Residency. (2014). *Library of Congress.* Retrievied from http://www.digitalpreservation.gov/ndsr/

[58] National Digital Stewardship Residency. (2014). *Harvard University*. Retrieved from http://projects.iq.harvard.edu/ndsr_boston/home

[59] National Digital Stewardship Residency. (2014). *Metropolitan New York Library Council*. Retrieved from http://ndsr.nycdigital.org/

[60] Organizations such as the Association for Library and Information Science Education (ALISE) are addressing areas such as the "Technology Competency Requirements of ALA-Accredited Library Science Programs," but the stewarding community may need more targeted research that attempts similar cross-organizational cataloguing of skills and competencies but with a more direct assessment of existing and needed digital stewardship skills. Scripps-Hoekstra, L., Carroll, M., & Fotis, T., (2013). Technology Competency Requirements of ALA-Accredited Library Science Programs: An Updated Analysis. *Articles.* Paper 45. Retrieved from http://scholarworks.gvsu.edu/library_sp/45/

nascent stages.[61] Cultural heritage institutions would benefit from an openness to recruiting highly specialized staff from non-traditional library disciplines to fill existing or newly-conceived positions in digital stewardship while providing training and education opportunities that enable digital stewards from across the professions to find commonalities.

**Actionable Recommendations**

- Continue to explore and expand models of support that provide interdisciplinary and practical experiences for emerging professionals and apply those models to programs for established professionals.
- Evaluate and articulate both the broad mix of roles and the specialized set of skills digital stewardship professionals are involved in.

*3.3 Multi-Institutional Collaboration*

It remains impractical for every institution to develop expertise in every aspect of the digital preservation challenge; different institutions could specialize in different aspects and rely on each other for some functions, spreading investments wisely where they might make a real impact. Transparency across organizations will avoid both duplication of effort and the over-reliance on single institutions by exposing organizational competencies and intentions. Transparency can then lead to optimal multi-institutional collaboration across a range of desired activities, including: fostering collaborative open source software development, sharing information on staffing and resources, engaging with standards and practice development, openly identifying stewardship responsibilities, and developing coordinate selection decisions and collection policies for born-digital acquisitions.

First, fostering community development of critical software infrastructure is critical to avoid catastrophic risks to stewardship infrastructure. The digital preservation community utilizes a mix of commercial, open source, project-funded, and homegrown standards, workflows, tools and services to perform digital stewardship tasks. Each organization selects the tools and services they need for the job based on their priorities, available funds and human resources. However, almost every institution active in digital preservation relies on some type of shared infrastructure—be it a specification or repository software. The support for these shared pieces of infrastructure also comes from a mix of non-profits, public and private libraries, service providers, and the open source community. However, overreliance on a single institution for supporting a piece of shared technical architecture can be risky.

For example, in the mid 2000s, in the international web archiving community, the International Internet Preservation Consortium,[62] in partnership with the Internet Archive, developed the core tools for harvesting the web at scale. The harvesting tool, Heritrix, and the access tool, Wayback, have been open source tools and are used by most of the major web archiving programs around the world. Their maintenance and development, however, was

[61] The federal government "Position Classification Standard for Librarian Series GS-1410" describes what it means to be a librarian in the Federal Service. This description for "librarian" hasn't been updated since 1994. Lazorchak, B. (2012) Is There a Future for Librarians? And Am I In It? *The Signal: Digital Preservation.* Retrieved from The Federal Library and Information Center Committee (FLICC) released an updated version of the "Federal Librarian Competencies" document in 2011 to define the knowledge, skills, and abilities (KSAs) needed to perform successfully as a federal librarian.

[62] http://netpreserve.org/

effectively solely supported by the Internet Archive. To provide more robust community development, over the last two years, the IIPC coordinated the re-launch of these shared tools, specifically Wayback,[63] to structure the tools as purely open source and to be sure the tool stays in-line with community needs and has a community "home." IIPC members are dependent on the shared infrastructure for web archiving and they put strategic effort toward supporting and maintaining the tools they rely on. The broader digital stewardship community needs to maintain awareness of its dependencies on shared infrastructure and tools, and to develop models for sustainable shared maintenance. A more detailed discussion of web archiving tools and the need for them to adapt to the modern web can be found in the Digital Content section below.

Second, information sharing is needed for efficient, sustainable development of preservation services. Institutions hosting repositories should be encouraged to document and publicly share their stewardship practices – even if they have no plans for "trusted" repository certification.[64] And the community, and standards organization should continue to develop light-weight methods for documenting and communicating stewardship practices, such as the NDSA Levels of Digital Preservation[65] and the Data Seal of Approval.[66]

If each institution cannot hire the required number of staff and variety of types of expertise, collaborative hiring and sharing of staff and skills could help. Developing robust community infrastructure requires making visible the different services offered, areas of expertise, and standards activities of organizations active in the digital preservation community. The community could then use that visibility to find opportunities where multiple organizations could benefit from a division of labor and identify gaps where something necessary is not getting done. This work would allow members in the community to identify potential specializations, and then to publicize commitments of organizations to specialize in a particular function so others can begin to rely on it. A key steps towards such information sharing, embodied by the NDSA Digital Preservation Staffing Survey,[67] is to identify preservation functions that could be outsourced, versus the functions that each organization prefers to (or must) do for itself—such as planning, alignment with parent organization's goals and designated communities. At the same time, it is essential that a market of preservation services develop so that organizations can supplement their in-house expertise with specialized services as needed, freeing organizations to staff in the areas most pertinent to their own competencies and resources.

Third, meaningful participation in the development of standards and policy outside of the cultural heritage institutions is critical to maintaining engagement with content users and providers. Standards and policies affecting how digital content is disseminated, retained, and used, are now being shaped in a variety of arenas. Many of these activities are focused on defining "open" and "public' access to data and information, rather than upon traditional records management and archiving. For example, current efforts through NISO to develop standards on *Open Access Metadata and Indicators*[68] and *Open Discovery*,[69] efforts by researcher and research

[63] https://github.com/iipc/openwayback
[64] A number of trusted repository certification audit tools have been developed, including one developed by MIT and hosted by Artefactual. See https://www.archivematica.org/wiki/Internal_audit_tool.
[65] http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/
[66] http://www.datasealofapproval.org/en/
[67] NDSA Digital Preservation Staffing Survey
http://www.digitalpreservation.gov/ndsa/documents/NDSA-staff-survey-poster-ipres2012.pdf
[68] http://www.niso.org/workrooms/oami/

to develop badges for data availability and reproducibility,[70] and commercial and government reexamination of information privacy practices,[71] all have the potential to dramatically affect the incentives for, content available to, and practice of digital stewardship. Ignoring such efforts is perilous as the NDSA report on recent PDF/A format standards makes clear—changes to the standard, which were incorporated largely without comment from the stewardship community, dramatically affects the durability of information in a format that was formerly a gold-standard for preservation: "The introduction of such a problematic new feature in the latest version of the PDF/A family suggests that perhaps the community of memory institutions need to take a more strategic, active, and vocal role in the standards development process." [72]

This high level of collaboration between many organizations requires several support elements to be in place.[73] The work that still needs to be done is at a community level and includes building organization's capacity to demonstrate trustworthiness, encouraging wide adoption of interoperability standards that would allow organizations to rely on each other more easily for predictable and equivalent outcomes, and establishing a method to ensure that digital preservation community interests are represented in all relevant standards bodies, and continuing to explore the benefits of certification and trust frameworks, including lightweight frameworks such as the Levels of Preservation.

In addition to articulating roles and sharing information about infrastructure, the community should openly identify stewardship responsibilities for specific kinds of content and share their acquisition priorities. More information is needed to evaluate how the digital stewardship community is doing in saving born-digital materials to which current and future researchers will want long-term access. Targeting collaborative efforts on evaluating national, regional, and local collecting priorities will be important first steps in understanding how to move toward a comprehensive and useful distributed collection of materials that provide illustration and evidence of our age. More specific recommendations on collecting and stewarding digital content are in section 4.

**Actionable Recommendations**

- Foster collaborative open source software development.
- Share information on staffing and resources.
- Engage with standards and practice development.
- Share selection decisions and collection priorities for born-digital acquisitions.

---

[69] Open Discovery Initiatives Working Group. (June 25, 2014). Open Discovery Initiative: Promoting Transparency in Discovery. NISO. Retrieved from http://www.niso.org/workrooms/odi/

[70] Grahe, Jon E. "Announcing Open Science Badges and Reaching for the Sky." *The Journal of Social Psychology* 154, no. 1 (2014): 1-3.

[71] The President's Council of Advisors on Science and Technology, *Big Data and Privacy: A Technology Perspective*, (2014). http://www.whitehouse.gov/sites/default/files/microsites/ostp/PCAST/pcast_big_data_and_privacy_-_may_2014.pdf

[72] Caroline Arms, et. Al. The Benefits And Risks Of The Pdf/A-3 File Format For Archival Institutions, (2014), NDSA Report. http://www.digitalpreservation.gov/documents/NDSA_PDF_A3_report_final022014.pdf

[73] See for example *Hess, C. ; Ostrom, E. Book Title: Understanding Knowledge as a Commons: From Theory to Practice Publisher : MIT Press;* Nancy McGovern & Katherine Skinner. 2012, Aligning National Approaches to Digital Preservation. Educopia Institute.

## 4. Technical Infrastructure Development

Broadly speaking, the infrastructure that enables digital preservation involves the staff, workflows, resources, equipment, and policies that ensure long-term access to digital information. This section focuses specifically on the technical component of that infrastructure. Technical infrastructure can be generally defined as the set of interconnected technical elements that provide a framework for supporting an entire structure of design, development, deployment, and documentation in service of applications, systems, and tools for digital preservation. This includes hardware, software, and systems. Organizational policies, practices, and regulations inform many of the observations and recommendations for the development of digital stewardship technical infrastructure.

*4.1 Coordinating an Ecosystem of Distributed Services*

The digital stewardship community has made great strides in identifying gaps in a modular, community-wide digital stewardship infrastructure and developing tools and services to fill those gaps. When NDIIPP started in 2000 there was very little digital preservation infrastructure in place. The Library of Congress worked with partners to identify gaps in digital stewardship infrastructure and directed resources to solutions to address infrastructure challenges that needed time to develop and that required a networked approach and broad community participation in order to be effective.

For example, Archive-It is a subscription web archiving service from the non-profit Internet Archive that helps organizations to harvest, build, and preserve collections of digital content. Portico is among the largest community-supported digital archives in the world and provides a sustainable economic model for how libraries, publishers, and funders can work together to preserve electronic journals, e-books, and other electronic scholarly content. The DuraCloud service addresses infrastructures to store and secure key digital content in a cost-effective manner across multiple cloud storage providers, both commercial and non-profit. The digital preservation community was eager to take advantage of cloud computing but were hampered by the challenges of dealing with multiple providers and understanding the technologies necessary to do digital preservation in the cloud. DuraCloud acts as an honest broker that understands the special requirements of preserving institutions.

In all three of these instances infrastructure challenges facing the digital preservation community were identified, and widely differing organizational structures proposed for supporting and stewarding a national collection of digital material of importance to the country. Still, challenges remain. At a broad level, a significant challenge for digital stewardship is the continued reliance on infrastructures subject to the vagaries of project-oriented funding and a lack of effective coordination in ensuring that community-developed tools have models for sustainability. These organizational challenges imply the ongoing need for some level of coordination across the infrastructure to ensure that effort is not duplicated and the development of authoritative resources that support coordination, like tool and format registries.

**Actionable Recommendation**

- Stewardship organizations and funders need to better identify and implement processes to maintain key software platforms, tools, and services. We need better models to support

long term sustainability for common goods.

*4.2 File Format Action Plan Development*

The sustainability of digital file formats and the risks of file format obsolescence persist as significant challenges for stewardship organizations.[74] Now that stewardship organizations are amassing large collections of digital materials, it is important to shift from more abstract considerations about file format obsolescence to develop actionable strategies for monitoring and mining information about the heterogeneous born-digital files the organizations are managing, especially the formats that don't result from digitization activities.[75] Recent studies of image formats and HTML doctypes[76] offers a valuable example of how organizations can analyze and share their data for analysis by third-party digital preservation researchers. By collecting and sharing this kind of data, it becomes possible for stewardship organizations to shift toward the development of file-format action plans based on the size of the risks that particular obsolescence threats pose to the significance those formats play in the organizations' managed content. Implementation of tools and services for creating file-format action plans is needed to make timely execution of file format plans a reality for data stewards.

While many organizations must be willing to accept any materials that they receive, file-format action plans suggest ways for organizations to prioritize resources towards digital formats with the greatest risk of obsolescence. At the same time, publicized format actions plans (such as the 2014 revisions by the National Archives and Records Administration[77]) drive practice by encouraging creating organizations to coalesce around a smaller set of possible digital format options, especially in industries with some degree of centralized control, such as federal, state, local and regional government. Going forward, it is critical that the format policies and action plans are translated into actions directly implemented and managed by tools and software. The implementation of file format policies in the Archivematica software platform illustrate a significant step forward in this effort[78].

It is also necessary for organizations to itemize and assess the digital content they are actively managing. For example, the Geospatial Data File Formats Reference Guide from the NDIIPP-supported Geospatial Multistate Archive and Preservation Partnership[79] project provides a quick reference[80] of some of the common geospatial raster and vector dataset types, and serves as a tool to quickly identify the geospatial file format types most commonly found in state government.

**Actionable Recommendations**

---

[74] Arms, Caroline & Fleischhauer, Carl. Digital Formats: Factors for Sustainability, Functionality, and Quality. IS&T Archiving 2005 Conference, Washington, D.C. http://memory.loc.gov/ammem/techdocs/digform/Formats_IST05_paper.pdf

[75] http://www.dlib.org/dlib/march14/rimkus/03rimkus.html

[76] Jackson, Andy. *Formats over Time: Exploring UK Web History.* http://arxiv.org/pdf/1210.1714v1.pdf 5 Oct 2012.

[77] http://blogs.archives.gov/records-express/2014/02/05/revised-format-guidance-issued/)

[78] Jordan, Mark. Automating the Preservation of Electronic Theses and Dissertations with Archivematica. 10th International Conference on the Preservation of Digital Objects, Lisbon, Portugal. http://summit.sfu.ca/item/13191

[79] http://www.geomapp.net/

[80] http://www.geomapp.net/docs/GeoMAPP_Geospatial_data_file_formats_FINAL_20110701.xls

- Stewardship organizations should document and share information about the file formats they currently manage to inform research and development.
- Stewardship organizations should prioritize the development of file format action plans that most appropriately reflect the kinds of content they are actually managing.
- To inform acquisitions, institutions should both make use of and comment on the newly released Library of Congress Recommended Format Specifications[81].

*4.3 Ensuring Content Integrity Across Infrastructure Migrations*

In 2011, when NDSA members were asked about plans for storage systems and architectures, 64% of respondents agreed or strongly agree that their organization planned to make significant changes in technologies in their preservation storage architecture within three years.[82] This underscores a fact that digital preservation practitioners already know quite well, that digital preservation is made possible through a long chain of migration through layers of current hardware and software systems to yet-to-be-established future infrastructures. This highlights the need for interoperability across different layers in these systems. In addition, easy migration of digital content from one system to another between organizations, such as vendor to client or partner to partner, would benefit the community enormously, particularly fostering the building of coalitions around preservation.

This points to a clear need for standards and the development of model plans for ensuring end-to-end data integrity in these migrations going forward. The key component of these standards and practices is tracking, maintaining, and auditing bit level fixity information. Much of the current practice is developed on an ad-hoc, one-off basis. Given that the forward cycle of migration will clearly be a continual part of digital preservation work, it is essential to develop clear guidance on how to plan for and manage these changes, and how to measure systematically the quality of the results. This kind of guidance development would inevitably point to issues that require further development of protocols and standards for interoperability and evaluation to help ensure continuity.

Also necessary, is the development of standards, practices, and strategies that directly address both lateral migration and forward migration. Case studies and more systematic reviews of activities currently underway need to be shared throughout the digital preservation community.

Fixity checking is of particular concern in ensuring content integrity. Abstract requirements for fixity checking can be useful as principals, but when applied universally can actually be detrimental to some digital preservation system architectures. The digital preservation community needs to establish best practices for fixity strategies for different system configurations. For example, if an organization were keeping multiple copies of material on magnetic tape and wanted to check fixity of content on a monthly basis, they might end up continuously reading their tape and thereby very rapidly push their tape systems to the limit of reads for the lifetime of the medium.

---

[81] Library of Congress. (2014-2015). Library of Congress Recommended Format Specifications. Retrieved from http://www.loc.gov/preservation/resources/rfs/rfs20142015.pdf

[82] Micah Altman, Jefferson Bailey, Karen Cariani, Michelle Gallinger, Jane Mandelbaum, Trevor Owens, 2013, NDSA Storage Report: Reflections on National Digital Stewardship Alliance Member Approaches to Preservation Storage Technologies. *D-Lib Magazine* 19(5/6)

There is a clear need for use-case driven examples of best practices for fixity in particular system designs and configurations established to meet particular preservation requirements. This would likely include description of fixity strategies for all spinning disk systems, largely tape-based systems, as well as hierarchical storage management systems. A chart documenting the benefits of fixity checks for certain kinds of digital preservation activities would bring clarity and offer guidance to the entire community. A document modeled after the NDSA Levels of Digital Preservation would be a particularly useful way to provide guidance and information about fixity checks based on storage systems in use, as well as other preservation choices.

**Actionable Recommendations**

- Support the development of standards, best practices and guidance for migrations.
- Support the development of best practices and guidance on fixity checking practices.

## 5. Research Priorities

This section focuses on areas of research that are critical to the advancement of both the basic understanding and the effective practice of digital preservation. Research in this area faces as funding challenge -- exemplified by Nobelist Eleanor Ostrom observation (to summarize her argument) that the fundamental and beneficial properties of knowledge in general and digital information specifically -- lead, somewhat paradoxically, to under-provisioning in any competitive market system.[83] To simplify -- knowledge is often easy to use without paying for it. Further, knowledge often yields its greatest benefits in the long term, making it difficult for institutions with short-term pressures to invest in. Thus it is perhaps no surprise that basic research in long-term access to knowledge -- digital preservation -- is under-resourced.

Cross-cutting issues of developing evidence-based and scalable curation practices are likely to remain strong priorities over the next seven years. And, in addition a number of areas for targeted research have emerged as critical for increasing the reliability and effectiveness of digital stewardship

Funding is needed in this area to develop basic theoretical models, extend the evidence base, and translate research findings into digital preservation practices and tools. Furthermore, digital preservation research is often closely tied to the development and evaluation of infrastructure, which makes it challenging to fund through basic research funding mechanisms. Decision-makers should recognize that basic research in these areas often needs to be paired with the development, support, and evaluation of infrastructure -- thus sustained support for infrastructure; and support for "action research" -- research-based practice; and for applied, practice-based research is called for.

### *5.1 Strengthening the Evidence Base for Digital Preservation*

A common challenge running through this report, and an overarching challenge for research is the limited amount of empirical evidence available. For example, this report makes

[83] E. Ostrom, and C. Hess, 2007, *Understanding knowledge as a commons: From theory to practice*." Massachusetts Institute of Technology

clear how effective digital preservation often requires answering questions such as: What content is already being effectively stewarded by other organizations? How much is the expected future cost of preserving that content? How often do different threats to information manifest: For example, what is the likelihood that: storage hardware or media fails; software errors cause information loss; stored information becomes inaccessible because of obsolete formats, or loss of other contextual knowledge; or that human error or maliciousness causes loss content in an information system? What is the reliability of current digital preservation networks and services? And how successful are other proposed strategies for replication, monitoring, certification, and auditing at preventing loss due to these threats?

The digital preservation community is beginning to develop a shared evidence base that can be used to answer these and similar questions. Recent medium-scale observational studies and field experiments have provided useful insights into the failure rates of spinning disk storage,[84] the proportion of files formats in use at a number of selected major digital repositories,[85] the long-term costs of preserving journal articles in pdf format,[86] and the extent and types of content being stewarded in institutional repositories.[87]

However, these studies must be broadened and repeated over time to establish a robust evidence base from which generalizable guidance can be drawn. Furthermore, for most questions in digital preservation, the current evidence base is constituted almost entirely of case studies. While case studies are useful for existence proofs, raising awareness of problems, process tracing, hypothesis generation, and other formative analysis, they are generally insufficient to advance our scientific knowledge, create robust predictive models, test causal hypotheses, or to strongly guide decision making.

For example, the NDIIPP program's highly informative case study/field experiment in the controlled transfer of complex collections of content demonstrated the challenges of content transfer and the likelihood of failures even in well-controlled cases.[88] However, to systemically guide decisions in this area, such case studies must be repeated longitudinally, repeated in different environments, and transformed, eventually into production public testbeds[89] and

---

[84] Pinheiro, E., Weber, W.D., & Barroso, L. A. (2007). Failure trends in a large disk drive population. In Proceedings of 5th USENIX Conference on File and Storage Technologies.

[85] Hitchcock, Steve, and David Tarrant. "Characterising and preserving digital repositories: File format profiles." Ariadne 66 (2011).

[86] Davies, Richard, et al. "How much does it cost? The LIFE Project-Costing Models for Digital Curation and Preservation." Liber Quarterly 17.3/4 (2007).

[87] Lynch, Clifford A., and Joan K. Lippincott. "Institutional repository deployment in the United States as of early 2005." D-lib Magazine 11.9 (2005): 5.; McDowell, Cat. "Evaluating institutional repository deployment in American academe since early 2005: Repositories by the numbers, part 2." D-lib Magazine 13.9 (2007): 3.

[88] Shirky, Clay. "Library of Congress Archive Ingest and Handling Test (AIHT) Final Report." NDIIPP. http://www. digitalpreservation. gov/partners/aiht/high/ndiipp_aiht_final_report. pdf (accessed April 22, 2011) (2005).

[89] Recent research by Becker, Faria, & Duretec, as part of the BenchmarkDP project, provides a potential model based framework for such testbeds.See: Christoph Becker, Luis Faria and Kresimir Duretec. Scalable Decision Support for Digital Preservation: An Assessment: OCLC Systems & Services, Emerald Publishing, 2015 (Forthcoming) and http://benchmark-dp.org/publications/

Becker, Christoph, and Kresimir Duretec. "Free benchmark corpora for preservation experiments: using model-driven engineering to generate data sets." In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries*, pp. 349-358. ACM, 2013.

conformance tests that can be used to rigorously compare approaches and systems. Furthermore shared, durable, community testbeds that provide a place where tools can be tried, common set of digital content to run trials provide a resource for systematically comparing proposing, incrementally improving practice, and calibrating both theory models and practical understanding.

Similarly, bit-level preservation, which is often characterized as one of the simpler, better understood areas of preservation, lacks systematic metrics and measurements for even simple failure scenarios.[90] Furthermore there is very little information on failures in complex systems using various redundancy, fixity, file transformation (compression, deduplication, encryption), auditing, and repair strategies.[91]

Moreover, a search of the discipline's key reference works, bibliographies, and literature databases reveal[92] very few rigorously validated preservation methods, wide-scale empirical studies, probability-based surveys or field experiments, replicable simulation experiments, public test corpuses, testbeds[93], or recognized conformance tests.[94] Although an applied field cannot rely on theoretical literature alone, it is essential to both grounded theory and robust practice that preservation strategies, methods, tools, and measures be formalized, standardized and evaluated systematically and rigorously. There are some datasets available that have been used by the digital stewardship community in the past, such as the Enron email dataset,[95] the September 11 Digital Archive , the Garfinkel, et al. GovDocs corpora[96], and the Geocities Special Collection 2009,[97] but more information is needed on how these datasets might be utilized in support of digital stewardship research. Broadly, across the field of digital preservation, there is an urgent need to develop a modular open and robust approach to testing, conformance, and measurement,

---

[90] See for an overview of current challenges: Rosenthal, David SH. "Bit preservation: a solved problem?." International Journal of Digital Curation 5.1 (2010): 134-148; and for a comprehensive model of bit preservation and approaches a comprehensive model. See: Zierau, Eld. "A Holistic Approach to Bit Preservation." (2011). Ph.D. Dissertation, Dept of Computer Science, University of Copenhagen

[91] Baker, M., Shah, M., Rosenthal, D. S. H., Roussopoulos, M., Maniatis, P., Giuli, T., et al. (2006). A fresh look at the reliability of long-term digital storage. In Proceedings of EuroSys2006.

[92] Borghoff, Uwe M. *Long term preservation of digital documents*. Springer, 2005.; Giaretta, David, 2011, Advanced Digital Preservation, Springer. Digital Curation Center, 2012, Curation Manual, http://www.dcc.ac.uk/resources/curation-reference-manual ; Bailey Jr, Charles W. "Digital curation bibliography: Preservation and stewardship of scholarly works." *Digital Curation Bibliography: Preservation and Stewardship of Scholarly Works* (2012). Force, Blue Ribbon Task. "Sustainable economics for a digital planet: Ensuring long-term access to digital information." *Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access* (2010).

[93] With notable, isolated exceptions such as Shirky, Clay. "Library of Congress Archive Ingest and Handling Test (AIHT) Final Report." NDIIPP. http://www. digitalpreservation.gov/partners/aiht/high/ndiipp_aiht_final_report. pdf (accessed April 22, 2011) (2005). and the Planets Testbed, Brian Aitken, Petra Helwig, Andrew Jackson, Andrew Lindley, Eleonora Nicchiarelli, Seamus Ross (2008). "The Planets Testbed: Science for Digital Preservation" , The Code4Lib Journal, Issue 3, 2008-06-23. Unfortunately, both of these projects have concluded and are no longer available.

[94] As noted above, current certifications are based primarily on process rather than demonstration of efficacy or outcome conformance.

[95] https://www.cs.cmu.edu/~enron/

[96] Garfinkel, Simson, Paul Farrell, Vassil Roussev, and George Dinolt. "Bringing science to digital forensics with standardized forensic corpora." digital investigation 6 (2009): S2-S11.

[97] https://archive.org/web/geocities.php

[98] in order to extend the evidence base on which preservation research and policy is founded.

Evidence is needed both to support either general selection of digital preservation practices and methods, or applications of selected digital preservation methods in a specific operational context. While preservation research should be better informed by cognate disciplines, research in information science and computer science generally target the functioning and use of information systems and are not focused on questions of long-term information access to, future understanding of, and value of steward digital collections that are central to preservation.

More resources are clearly needed for research—especially in the US. While there are a number of major research efforts in support of digital preservation, much of it is in other countries. The Alliance for Permanent Access[99] (APA), in Europe, has been an important incubator for digital preservation research in trust, sustainability, usability, and access. The SCAPE[100] project, funded by the EU, is advancing the evidence base on format migration, and format risks, repository performance, and collecting and sharing evidence generally. The 4C[101] project, also funded by the EU is conducting much-needed research in the cost and economic modeling of preservation. Complementary efforts in the US are needed. Further, most of the EU projects described are funded on fixed terms (some already completed) -- which makes the development of sustainable tools and testbeds challenging.

What is also needed is to apply the research methodologies already used in other fields that rely heavily on observation of human and system behavior. This includes methodologies such as: probability-based surveys of information management practice and outcomes; replicable simulation experiments, and theoretically grounded new practices, tools, and methods; and field experiments, in which randomized interventions are applied and evaluated in real operational environments.

**Actionable Recommendations**

- Funders should give priority to programs that systematically contribute to the overall cumulative evidence base for digital preservation practice and resulting outcomes -- including supporting testbeds for systematic comparison of preservation practices
- Funders should give priority to programs that rigorously integrate research and practice
- Research based evaluation of practice should go beyond case-studies in their approach, and include replicable methods to support systematic inference

*5.2 Stewardship at Scale*

A second cross-cutting research problem is dealing with scale in digital stewardship. Digital collections are growing exponentially. Keeping track of everything and being able to work with and manage content is increasingly difficult. Growing volumes of digital materials will test the financial and operational capabilities of organizations engaged in preservation activities. Of particular concern are issues around the stewardship of big data and the search and

[98] For a possible approach see, Becker, Christoph, and Andreas Rauber. "Decision criteria in digital preservation: What to measure and how." *Journal of the American Society for Information Science and Technology* 62.6 (2011): 1009-1028.

[99] http://www.alliancepermanentaccess.org/

[100] http://www.scape-project.eu/

[101] http://www.4cproject.eu/

indexing of digital collections at scale.

*5.2.2 Stewardship of "Big" Collections*

Institutional responsibilities to serve and preserve big data will also be influenced by user and content creator expectations regarding its maintenance and accessibility. Storage, intellectual and administrative control, and access will all be redefined by the demands of big data. Currently, many organizations lack the expertise or economies-of-scale to store and process petabytes of data.

"Big" data can create scaling challenges not only as a result of pure volume, but for other reasons, including the numbers of objects that must be curated, the velocity (frequency) with which data objects and collections are updated, and the variety (heterogeneity) of the data objects, formats, and characteristics. Thus scaling challenges go far beyond the bare provisioning of storage—with variety often being the biggest challenge for institutions.[102] Scaling to billions of files, and/or to individual files of extremely large size, renders manual methods of archival selection, quality- evaluation and control all but impossible, creates performance challenges for data ingestion workflows and tools increases the complexity of indexing and discovery, and may render standard computer-human interfaces used for curation and user access unusable. Further, we are just beginning to understand how the scale of "big data" affects privacy, confidentiality, and personally-identifiable information, and the implications that this has for managing such data in the future.

Moreover, scaling collections presents special challenges for data stewardship and long-term access: The increased number and size of files, and increased volume of collections, can overwhelm the standard strategies for replication, fixity checking, and repair that are needed to ensure long-term data integrity.[103] Increasing the number of formats and objects types creates challenges for the in-depth documentation, format characterization, and format migration that are required to maintain long-term accessibility. Increasing velocity of data creates particular challenges for maintaining the versioning and provenance required of durable, authentic collections.

This lack of infrastructure and expertise will require collaborative solutions involving greater automation, scalable processes, and modular, adaptive frameworks. Community-driven scalable solutions to a wide-range of unique and independent preservation activities must be developed. In addition, the establishment of shared infrastructure and open-source solutions will enable greater efficiency and economic feasibility towards the growing volume of digital content that must be preserved.

There are persistent issues in terms of indexing and searching across large amounts of content, especially while ensuring moderate reads on content stored on magnetic tape. It is no longer enough to rely on increasingly expensive and fast drives and systems. At this point, there are opportunities to exploit efficiencies in the design of smarter systems and architectures. For example, one might rebuild parts of an index when an error occurs to avoid having to restage the full index.

In sum, there is a need for collaborations with other groups and initiatives in fields

[102] K. De Souza, Realizing the Promise of Big Data: Implementing Big Data Projects (2014), Report. IBM Center for the Business of Government.

[103] Rosenthal, David SH. "Bit preservation: a solved problem?." International Journal of Digital Curation 5.1 (2010): 134-148.

addressing issues of scale in digital data, to work on common use cases and to optimize opportunities for building or acquiring cost-effective common solutions.

*5.2.2 Developing Systematic Value Models for Selection at Scale*

It is neither desirable nor feasible to keep all research information forever – thus selection and appraisal are critical part of data curation.[104] However, estimating the value of information is inherently difficult. Arrow's information paradox states that ex-ante a buyer cannot assess the value of particular information – it can only be known ex-post, at which point the buyer has limited incentive to pay for it.[105]Although assignment of intellectual property rights can address this issue to a limited extent, it is very challenging[106] – and hence markets for information goods are generally thin. And although "data quality" is sometimes seen as a proxy for value, no feasible universal quality measure exists -- data quality measures are notoriously varied, discipline specific, contextual, and difficult to implement in practice.[107] Furthermore, intellectual property rights notwithstanding, the non-consumptive and limited excludability that are inherent properties of information goods implies that any pure market solution will produce and distribute information at levels that are socially sub-optimal. (Hess & Ostrom 2006) Moreover, the future value of research information and its communication potential are notoriously difficult.

The development of economic models, methods, and empirical analysis that would lead to more rigorous, reliable, and systematic evaluation of the value of research information constitutes an important, but poorly understood, set of problems. Researchers and curators continually make implicit or explicit decisions regarding what information to retain, how long to retain it, what effort to expend in making it accessible and understandable, and when that effort should be applied.

Correctly estimating the future value of a single specific information object or collection is often impossible or impractical -- similar to trying to guess the future stock price of a single corporation. Estimating the value of portfolios, however, is standard practice in finance, and could become standard practice in digital curation.

Two promising areas to explore in this pilot are economic portfolio theory and information science threat taxonomies. Historically, selection criteria have been made locally, and in an ad-hoc manner, based on the history and local values of the institution selecting. In economics research generally contingent valuation surveys[108] are a standard tool for measuring the value of non-market goods – yet this method has never, to our knowledge been applied to valuing research data. Similarly portfolio selection modeling[109] is the primary tool used in economics to diversify across risky investments, but has never been applied to research data. The Center will import these economic research methods to the study of research data. Diversification is also an essential strategy for mitigating risks to future access. There is a

[104] Giaretta, D. (2011), Advanced Digital Preservation, Springer.
[105] Arrow, Kenneth J. "The value of and demand for information." Decision and organisation. Londres: North-Holland (1972).
[106] Gans, J. S., & Stern, S. (2010). Is there a market for ideas?. Industrial and Corporate Change, 19(3), 805-837.
[107] Altman, M. (2012). "Mitigating Threats To Data Quality Throughout the Curation Lifecycle. In G. Marciano, C. Lee, & H. Bowden (Eds.), Curating For Quality (pp. 1–119). Retrieved from http://datacuration.web.unc.edu/
[108] Mitchell, R. C., & Carson, R. T. (1989). Using surveys to value public goods: the contingent valuation method. Rff Press.
[109] Markowitz, H.M. (March 1952). "Portfolio Selection". The Journal of Finance 7 (1): 77-91

well−identified taxonomy of potential single−points−of−failure (highly correlated risks), that at minimum, a trustworthy preservation system should mitigate: These risks include media failure, hardware failure, software failure, communication errors, network failure, media and hardware obsolescence, software obsolescence, operator error, natural disaster, external attack, internal attack, economic failure, and organizational failure.[110] Nonetheless, the reliability, design, and behavior of both centralized and distributed preservation networks is just beginning to be understood. A notable exception is Baker, *et* al.[111] which employs Monte-Carlo simulation to explore trade-offs in costs and reliability across bit-level replication technology choices. Designing effective technical diversification strategies for long-term access requires more extensive modeling along these lines.

**Actionable Recommendations**

- Funders and researchers should prioritize programs and projects that increase the scalability of digital stewardship.
- Researchers should recognize that the challenges of "big" collections goes beyond size and storage,, to dealing with the variety, and velocity of big data and big collections across all phases of the curation lifecycle.
- Researchers and funders should recognize that selection and appraisal is a fundamental challenge at scale -- and prioritize systematic, evidence based, non-labor intensive methods of evaluating portfolios of information.

### *5.3 Targeted Applied Research Areas*

A number of research issues, are less universal than those of scale and evidence, but are vital in order to develop more effective, reliable, and efficient tools, models, and methods for digital stewardship in the next three to five years.

#### *5.3.1 Applied Research for Cost Modeling*

In the near term, there are specific areas of applied research around digital preservation lifecycle issues that need attention. Currently there are limited models for cost estimation for ongoing storage of digital content; cost estimation models need to be robust and flexible. There are bodies of written research on the topic that explore the costs of specific use cases,[112] and Lavoie & Grindley,[113] as part of the 4C project and building on the Blue Ribbon Task Force report, have developed a high-level conceptual framework for economic sustainability that outlines major lifecycle phases, stakeholders, economic conditions, and risks relevant to digital

[110] Reich, V., Rosenthal, D. S. H., Robertson, T., Lipkis, T., & Morabito, S. (2005). Requirements for Digital Preservation Systems: A Bottom-Up Approach. D-Lib Magazine, 11(11).
[111] Baker, M., et al. (2006) A fresh look at the reliability of long-term digital storage, Proceedings of the 1st ACM SIGOPS/EuroSys European Conference on Computer Systems. ACM; 2006.; p. 221-234.
[112] http://blogs.loc.gov/digitalpreservation/2014/01/a-national-agenda-bibliography-for-digital-asset-sustainability-and-preservation-cost-modeling/
[113] B. Lavoie & K. Grindley, "Draft Economic Sustainability Reference Model" (2013) 4C Project, http://www.4cproject.eu/community-resources/outputs-and-deliverables/ms9-draft-economic-sustainability-reference-model

preservation. However, there is still a need to more clearly define the staffing aspects of digital stewardship cost (as in the staffing surveys mentioned earlier in the Organizational Roles, Policies and Practices section) and to develop models that systematically and reliably predict the future value of preserved content. Furthermore, many long-term cost models are based on assumptions that the historical rate of decrease in storage prices will continue indefinitely—an assumption that is contradicted by a careful analysis of cloud storage trends and emerging storage technologies costs.[114]

Different approaches to cost estimation should be explored and compared to existing models with emphasis on reproducibility of results. The development of a cost calculator would benefit organizations in making estimates of the long-term storage costs for their digital content.

Further, as discussed in other sections, there are many opportunities to develop better value models and business models: In *Developing Systematic Value Models for Selection at Scale* we discuss the challenges of systematically and reliably predicting the future value of portfolios of preserved content. In we discuss new form of business models & business opportunities, many of them collaborative. A combination of value, cost, and business model development is needed for rational and efficient digital curation

This research needs to address multiple storage models: Locally stored data, distributed preservation networks, data cooperatives, cloud storage, brokered cloud storage systems and hybrid systems should each be addressed in cost models s so that organizations can make informed cost-effective digital preservation decisions.

Further, as discussed in other sections, there are many opportunities to develop better value models and business models, many of them collaborative. A combination of value, cost, and business model development is needed for rational and efficient digital curation

### *5.3.2 Environmental Sustainability and Sustainability of Digital Collections*

As our digital cultural and scientific heritage grows at an exponential rate, it is often easy to overlook the underpinning material costs. Data, of course, are not "virtual" or "ephemeral"; rather, every byte requires resources to ensure its reliable storage and accessibility. Recent reports suggest that data management currently taxes upwards of 2% of total global energy consumption.[115] There is a growing body of work on low-carbon "green" computing and data centers.[116] However, there is currently no substantive body of work connecting this work on environmental sustainability to economic modeling for long-term digital storage.

---

[114] Rosenthal, David SH, Daniel C. Rosenthal, Ethan L. Miller, Ian F. Adams, Mark W. Storer, and Erez Zadok. "The economics of long-term digital storage." Memory of the World in the Digital Age, Vancouver, BC (2012).

[115] Glanz, James. "Data Centers Waste Vast Amounts of Energy, Belying Industry Image." The New York Times, September 22, 2012, sec. Technology. http://www.nytimes.com/2012/09/23/technology/data-centers-waste-vast-amounts-of-energy-belying-industry-image.html.

[116] Climate Group, The. SMART 2020: Enabling the Low Carbon Economy in the Information Age. Global eSustainability Initiative, 2008. http://www.smart2020.org/_assets/files/02_Smart2020Report.pdf; Cook, Gary. How Clean Is Your Cloud? Greenpeace, April 2012. http://www.greenpeace.org/international/Global/international/publications/climate/2012/iCoal/HowCleanisYourCloud.pdf; Google's Green Data Centers: Network POP Case Study. http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en/us/corporate/datacenter/dc-best-practices-google.pdf; Masanet, Eric, Arman Shehabi, and Jonathan Koomey. "Characteristics of Low-carbon Data Centres." Nature Climate Change 3, no. 7 (July 2013): 627–630. doi:10.1038/nclimate1786.

A comprehensive examination of digital environmental sustainability requires an interdisciplinary perspective that merges material and access needs, and brings together digital preservationists, IT specialists, administrators, computer engineers, and environmentalists. Metrics by which to evaluate the operational costs of data centers, such as those produced by the non-profit organization Green Grid, or the JISC-funded Greening Information Management Assessment Framework, offer ways in which digital preservationists may conduct preliminary, quantifiable assessments. These are only first steps, however, and a much more comprehensive, interdisciplinary approach is needed that takes into account issues of digital stewardship. There is a need for basic research and development, in particular new case studies that could refine current metrics, as well as a need to investigate ways of educating the broader community about sustainability.

### *5.3.3 Computable Information Equivalence & Significant Properties*

Long term management of digital content almost always involves changing the representation of that content while retaining its meaning. Thus the "significant properties"[117] of content -- identifying the properties of that content that give it meaning -- has emerged as a key concept in digital preservation, and generated a focused and influential body of research. [118] The concept of significant properties can be applied to all content types -- and recent expansions focus on priority content areas such as software and data[119], that are discussed in the *Content* section above. Moreover, this line of research has implications across a diverse set of applications including format selection and migration; quality measurement and control; rights management; and information discovery and retrieval.

Research into computing significant properties and creating semantic fingerprints or is now supporting innovative preservation practices such as ensuring integrity across format migrations in the Dataverse Network system;[120] for quality assurance in audio preservation in the SCAPE project[121]. As exciting, the technology of this semantic fingerprints is rapidly developing, and its use widespread in the commercial sector -- where numerous consumer services such as SoundCloud enable tens of millions of users to identify music base on quick, wide-scale audio fingerprinting; and video services such as YouTube use video fingerprinting as a core part of their digital rights management and ad placement process.

Although widely used in the commercial sector, methods for scalable evaluation of semantic similarity is far less common in digital preservation practice. Yet, the multiplicity of instantiations of the same or similar digital objects illustrates the need for and application of basic research to explore the many ways multiple digital objects could contain equitant informational content given different contexts of significance. For instance, a single photograph may be represented by any number of derivative files of varying sizes, in varying formats, and

---

[117] This term was first coined in Hedstrom, Margaret, and Christopher A. Lee. "Significant properties of digital objects: definitions, applications, implications." In *Proceedings of the DLM-Forum*, pp. 218-27. 2002.
[118] See Giaretta, David, 2011, Advanced Digital Preservation, Springer
[119] "A Framework for Applying the Concept of Significant Properties to Datasets." *Proceedings of the 74th Annual Meeting of the American Society for Information in Science and Scholarship*. Simone Sacchi, Karen M Wickett, David Dubin, Allen H. Renear (2011).
[120] Crosas, Merce,. "The dataverse network®: an open-source application for sharing, discovering and preserving data." D-lib Magazine 17, no. 1/2 (2011).
[121] Jurik, Bolette Ammitzbøll, and Jesper Sindahl Nielsen. "Audio quality assurance: An application of cross correlation." *Proceedings of IPres* (2012): 144-149.

with different sets of embedded metadata inside it.[122] Similarly, an organization may have 15 PDFs of the same article each with a different cover page, but all of which are substantively identical. Preservation research needs to map out the networks of similarity and equivalence across different instantiations of objects so that they can make better decisions on how to manage content, bearing in mind what properties of a given set of digital objects are significant[123] to their particular community of use. Research is also required in order to characterize quality and fidelity dimensions and create methods for computing format-independent fingerprints of content,[124] so that the fidelity of digital objects can be effectively managed over time. In this space, there is potential value in creating semantic fingerprints through fuzzy hashing algorithms that can map out the similarity of bitstreams, applications to analyze and compare rendered content in different formats (image comparison, extracting and comparing sound frequencies across audio and video files, etc.), and other innovative potential modes for asserting that some aspect of a given set of objects is similar in a particular way to another set of objects. Beyond basic research to develop methods for identifying information equivalence, there is a need for research in different usage contexts to understand when particular modes or levels of information equivalence are relevant to particular stakeholders in particular contexts.

*5.3.4 Policy Research on Trust Frameworks*

There is a well-identified taxonomy of potential single-points-of-failure (highly correlated risks), that at minimum, a trustworthy preservation system should mitigate: These risks include media failure, hardware failure, software failure, communication errors, network failure, media and hardware obsolescence, software obsolescence, operator error, natural disaster, external attack, internal attack, economic failure, and organizational failure.[125]

Geographic risk, curatorial error, internal malfeasance, economic failure, and organizational failure require that replications be diversified across distributed, and often, collaborative organizations.[126] No one provider can or should provide all elements of long-term preservation—therefore developing approaches for collaborative stewardship and for modularized review, auditing, and certification is required.

In this area, community use of collaborative institutional mechanisms to mitigate preservation risk is growing. This is reflected in the growth of organizations such as the Global LOCKSS Network, Data-PASS, MetaArchive, the Digital Preservation Federation, and the Digital Preservation Network. These organizations, and the multi-institutional stewardship

---

[122] C. Marshall Digital Copies and a Distributed Notion of Reference in Personal Archives in *Digital Media: Technological and Social Challenges of the Interactive World* (2011) edited by W Aspray, M Winget.

[123] Hedstrom, Margaret, and Christopher A. Lee. "Significant properties of digital objects: definitions, applications, implications." *Proceedings of the DLM-Forum*. 2002.

[124] Altman, Micah. "A fingerprint method for scientific data verification." *Advances in Computer and Information Sciences and Engineering*. Springer Netherlands, 2008. 311-316.

[125] D. S. H. Rosenthal, T. Robertson, T. Lipkis, V. Reich, and S. Morabito, "Requirements for Digital Preservation Systems", D-Lib Magazine 11 (11), 2005 http://www.dlib.org/dlib/november05/rosenthal/11rosenthal.html

[126] Altman, M., Beecher, B., & Crabtree, J. (2009). A Prototype Platform for Policy-Based Archival Replication. Against the Grain, 21(2), 44-47. http://www.box.net/shared/gxdcnsxunlpg9xol5h1t
Altman, M., & Crabtree, J. (2011). Using the SafeArchive System : TRAC-Based Auditing of LOCKSS. Archiving 2011 (pp. 165-170). Society for Imaging Science and Technology.
http://www.box.net/shared/8py6vl9kxivo6u21rkn8

approach they represent, have increased both in use and in recognition.

Nonetheless, the reliability, design and behavior of both centralized and distributed preservation networks are just beginning to be understood. It is critical to develop robust trust frameworks that address these risks, because institutions need to be able to measure, evaluate, and monitor the reliability and trustworthiness of trustworthy repositories, collaborating organizations, and third-party services (such as cloud computing). Measuring and evaluating the trustworthiness of such organizations and services is a substantial challenge for policy research.

The preservation community has made progress in this direction: Many of the processes identified with trustworthy content stewardship have been recognized, standardized, and documented in the Trustworthy Repositories Audit & Certification (TRAC) criteria and in the newly released ISO standard that has succeeded it. Furthermore a number of organizations are moving toward formal certification under this ISO 16363 standard.

Notwithstanding, much remains to be done. Few stewardship organizations have obtained trustworthy certification, and a relatively small percentage of stewardship organizations are seeking it,[127] while many third-party services will not seek it at all. Furthermore, no certification process has yet been widely recognized by the preservation community.

A number of approaches to certification auditing and assessment are particularly promising. Self-evaluation of trusted repository criteria, complemented by peer review and other community-based assessment may be both more reliable and less burdensome that external certification. Modularization of assessment and auditing, in which audits apply to particular subsets of criteria and responsibilities, is another promising approach. Some examples of work in this area include the *Data Seal of Approval,* which relies on peer review of self-assessments, and the NSA Standards Group ongoing project on understanding options for addressing standards and requirements..

The current trusted repository approach relies upon a very small subset of mechanisms employed in trust engineering. The role of many other approaches are both underemployed and poorly understood. *Moreover, reliability, effectiveness, and costs of current trust frameworks, including TRAC and ISO16362 has yet to be empirically demonstrated and systematically measured: How reliable are certification procedures, self-evaluations, and the like at identifying good practices? How much do the implementations of such practices actually reduce risk of loss? The evidence base is not yet rich enough to answer these questions.*

In terms of audit modeling, there is now more discussion on issues of file fixity and authenticity in the digital stewardship community,[128] but the discussions are still at a mostly introductory level. The NDSA is developing a report on file fixity issues and this may encourage research paths. The stewardship community should leverage the work being done in computer science on information security and file auditing to avoid reinventing the wheel.

For example, transparency is another key mechanism for mitigation of the risks above, but it is currently underutilized and poorly understood within this domain. Implementation transparency implies the use of open protocols, and often implies the use of open source (or protocols and algorithms with independent, open implementations). Operational transparency involves demonstrating both the process and the evidence that enables others to independently verify that services are being met. The NDSA principles are an example of a generally stated

[127] Altman, et. al "Reflections on National Digital Stewardship Alliance Member Approaches to Preservation Storage Technologies" 2013, *D-Lib.*

[128] http://blogs.loc.gov/digitalpreservation/2014/02/check-yourself-how-and-when-to-check-fixity/

organizational transparency goal, although these fall short of being enforceable.
It is particularly important that high-level policies such as TRAC can be demonstrated at the level of operations and systems action. Systematic auditing is necessary to keep cloud storage honest, and Shah et al. (2007) provide some guiding technical principles for developing related auditing mechanisms.[129] Work such as project Pledge, the SafeArchive system, and iRods have demonstrated that complex policies can be successfully mapped to systems behavior and transparently audited.[130] The SafeArchive[131] system and other bit-level auditing practices could be connected to the NDSA Levels of Preservation[132] work to help organizations determine and validate the costs of scaling different auditing schemes.

In general, further research is needed in the design, implementation, and evaluation of trustworthy digital stewardship mechanisms and their use,[133] including: building an organization's capacity to demonstrate trustworthiness, rewards, and penalties; peer review; statistical quality control and reliability estimation; incentive compatible mechanisms; threat-modeling and vulnerability assessment; portfolio diversification models; transparency and the release of information permitting direct evaluation of compliance; cryptographic approaches, including cryptographic signatures over semantic content; and generating and managing social evidence of compliance.

**Actionable Recommendations**

- Funders and researchers should prioritize a number of targeted applied areas of research that constitute special opportunities for improving the reliability and efficiency of preservation practice, including: cost modeling, environmental sustainability, computable significant properties, and trusted frameworks for stewardship

[129] Shah, Mehul A., Mary Baker, Jeffrey C. Mogul, and Ram Swaminathan. "Auditing to Keep Online Storage Services Honest." In *HotOS*. 2007.

[130] Altman & Crabtree 2012. Smith, MacKenzie, and Reagan W. Moore. "Digital archive policies and trusted digital repositories." (2007).

[131] www.safearchive.org

[132] http://blogs.loc.gov/digitalpreservation/2012/11/ndsa-levels-of-digital-preservation-release-candidate-one/

[133] See B. Schneier, 2012. Liars and Outliers, John Wiley & Sons for a review of trust engineering approaches.

**Acknowledgements**
The NDSA would like to thank the reviewers and initial readers of the *2015 Agenda* for their thoughtful comments which greatly improved the document.

**About the NDSA**

Founded in 2010, the National Digital Stewardship Alliance (NDSA) is a consortium of institutions that are committed to the long-term preservation of digital information. NDSA's mission is to establish, maintain, and advance the capacity to preserve our nation's digital resources for the benefit of present and future generations. The NDSA comprises over 160 participating institutional members. These members come from 45 states and include universities, consortia, professional societies, commercial businesses, professional associations, and government agencies at the federal, state, and local level. NDSA organizations have proven themselves committed to long-term preservation of digital information.

To learn more about the NDSA: http://www.digitalpreservation.gov/ndsa/

**Be a Part of the Conversation**

Comments welcome at ndsa@loc.gov or @NDSA2.

**About the Authors**

The leadership group of the NDSA authored this report and engaged in discussions with the NDSA membership to identify significant trends and challenges in digital stewardship. The membership of the NDSA contributed markedly to these discussions. The leadership group is made up of the Coordinating Committee members, the Working Group co-chairs, and the NDSA facilitator:

**Micah Altman**
*Director of Research, MIT Libraries, MIT; Non-Resident Senior Fellow, Brookings Institution.*
Dr Micah Altman is Director of Research and Head/Scientist, Program on Information Science for the MIT Libraries, at the Massachusetts Institute of Technology. Dr. Altman is also a Non-Resident Senior Fellow at The Brookings Institution. Dr. Altman conducts research in social science, information science and research methods—focusing on the intersections of information, technology, privacy, and politics, and on the dissemination, preservation, reliability and governance of scientific knowledge. Altman serves on the NSDA Coordinating Committee.

**Jefferson Bailey**
*Partner Specialist & Program Manager, Internet Archive*
Jefferson Bailey works on Archive-It and other programs involving web archiving partnerships with the library, archive, museum, and education communities. He was formerly the Strategic Initiatives Manager at Metropolitan New York Library Council (METRO), a Fellow in Digital

Preservation at the Library of Congress, worked on digital projects at Brooklyn Public Library and Frick Art Reference Library, and has done archival work at NARA, NASA, and The New York Times. He is co-chair of the NDSA Innovation Working Group, on the Steering Committee of the International Internet Preservation Consortium (IIPC) and is a visiting lecturer in University of Pittsburgh's Library & Information Science program.

**Karen Cariani**
*Director of Media Library and Archives, WGBH*
Karen Cariani has worked at WGBH since 1984 in television production and archival-related roles. She has 20-plus years of production and project management experience, has worked on numerous award-winning historical documentaries, and has been project director for many critical projects. She worked with the WNET, PBS, NYU and WGBH Preserving Public Television partnership as part of the Library of Congress National Digital Information Infrastructure Preservation Project. She served two terms (2001-2005) on the Board of Directors of Association of Moving Image Archivists (AMIA). She also serves on Digital Commonwealth executive committee. Recent projects include managing the American Archive Inventory project for CPB, and project director of PBCore development and Boston Local TV News Digital Library project. Cariani is the co-chair of the NDSA Infrastructure Working Group.

**Jim Corridan**
*State Archivist and Director of the Indiana Commission on Public Records*
Jim Corridan was elected to the Board of Directors of the Council of State Archivists (CoSA) in 2009, where he helped establish and chaired CoSA's State Electronic Records Initiative (SERI) in 2011 and currently serves as President of CoSA. SERI is focused on governance, best practices, awareness, and education to strengthen all the state and territorial archives in the United States. Working with the staff from the Library of Congress, Corridan and the Indiana State Archives hosted the first regional Digital Preservation Outreach and Education (DPOE) project, the Midwest train-the-trainer program, in August of 2012 and has been an advocate for digital preservation. He is State Archivist of Indiana.

**Jonathan Crabtree**
*Assistant Director of Computing and Archival Research at the Odum Institute for Research in Social Science at UNC Chapel Hill*
As assistant director, Jonathan Crabtree completely revamped the institute's technology infrastructure and has positioned the institute to assume a leading national role in information archiving. Crabtree's 22 years of experience in information technology and networking as well as his engineering background bring a different perspective to his current role. Crabtree serves on the NSDA Coordinating Committee.

**Michelle Gallinger**
*Information Technology Specialist, Library of Congress*
Michelle Gallinger works to develop digital preservation communities. Gallinger develops policies and guidelines for digital preservation practices, life-cycle management of digital materials, and stakeholder engagement at the Library of Congress. She also provides strategic

planning for the National Digital Information Infrastructure and Preservation Program.

**Andrea Goethals**
*Digital Preservation and Repository Services Manager, Harvard Library*
Andrea Goethals is responsible for providing leadership in the development and operation of Harvard's digital preservation program and for the management and oversight of the Digital Repository Service (DRS), the university's large-scale digital preservation repository. She is currently involved in the rolling out of the next-generation DRS ("DRS2"), several pilot projects (email archiving, exposing public APIs to access DRS content and metadata), and planning for new digital preservation services at the university. She is on the Curriculum Committee of the National Digital Stewardship Residency program and participates in the International Internet Preservation Consortium's Preservation Working Group. She is the co-chair of the NDSA Standards and Practices Working Group.

**Abbie Grotke**
*Lead Information Technology Specialist, Library of Congress*
Abbie Grotke is the Web Archiving Team Lead at the Library of Congress and a member of the NDIIPP team. She came to the Library in 1997 to work on American Memory digitization projects. Since 2002 she has been involved in web archiving and the digital preservation program at the Library of Congress. Grotke currently serves on the Steering Committee of the International Internet Preservation Consortium and is the co-chair of the NDSA Content Working Group.

**Cathy Hartman**
*Associate Dean of Libraries, University of North Texas*
Cathy Hartman's preservation efforts began in 1997 by establishing the CyberCemetery to preserve the Websites of U.S. government agencies and commissions and creating partnerships at Federal and State levels to archive electronic government information. She founded the Portal to Texas History to enable collaborations with more than 200 libraries and museums to digitize and preserve their collections for Web access. She currently chairs the International Internet Preservation Consortium Steering Committee and co-chairs the NDSA Content Working Group.

**Butch Lazorchak**
*Information Technology Specialist - Project Manager, Library of Congress*
Butch Lazorchak is a digital archivist in the National Digital Information Infrastructure and Preservation Program at the Library of Congress. He sits on the NDIIPP Communications Team, working to expand awareness of digital preservation issues and opportunities through a variety of channels, including regular postings for The Signal blog. He is the co-chair of the NDSA Outreach Working Group.

**Jane Mandelbaum**
*Manager of Special Projects, Office of the Director for Information Technology Services, Library of Congress*
Jane Mandelbaum is currently leading and guiding enterprise-wide projects and architecture initiatives for large-scale, high-performance digital storage and archiving. She previously served

as IT implementation and operations manager for a number of large IT systems at LC and led a team to establish and operate the Library's end-user computing environment. Mandelbaum is the co-chair of the NDSA Innovation Working Group.

**Carol Minton Morris**
*Director of Marketing and Communications, DuraSpace*
Minton Morris (Terrizzi) joined the National Science Digital Library (NSDL) team at Cornell University in 2000 and served as Communications Director from 2000-2009. She was the Communications Director for the Fedora Commons organization from 2007-2009. As Director of Marketing and Communications for the DuraSpace organization since 2009 she leads strategic editorial content and materials planning, development and distribution focused on sustaining open source projects (DSpace, Fedora) and marketing services (DuraCloud, DSpaceDirect). She is deputy co-chair of the annual International Open Repositories Conference and serves as co-chair of the NDSA Outreach Working Group.

**Kate Murray**
*Information Technology Specialist (Audio-Visual Specialist), Library of Congress*
At the Library of Congress, Murray primarily works with the Federal Agencies Digitization Guidelines Initiative (FADGI) and on the Sustainability of Digital Formats website. Prior to joining the Library of Congress, Murray was a Digital Process Development Specialist in the Digitization Planning Branch at the National Archives and Records Administration (NARA) specializing in standardizing and documenting moving image and audio formats. She is the immediate-past co-chair of the Association of Moving Image Archivists (AMIA) Preservation Committee and is a member of SMPTE, ARSC and AES. She is the co-chair of the NDSA Standards and Practices Working Group.

**Trevor Owens**
*Digital Archivist, Library of Congress*
At the Library of Congress, Trevor Owens works on the open source Viewshare cultural heritage collection visualization tool, as a member of the communications team, and on convening relevant stakeholders about preservation issues. Before joining the Library of Congress he worked for the Center for History and New Media and before that managed outreach for the Games, Learning, and Society Conference. Owens is the co-chair of the NDSA Infrastructure Working Group.

**Megan Phillips**
*Electronic Records Lifecycle Coordinator, National Archives and Records Administration*
Megan Phillips assists senior management with Electronic Records Archives (ERA) planning and the coordination of electronic records units and projects. In addition, Phillips is responsible for gathering and helping NARA prioritize the business requirements for the development and evolution of ERA. Phillips serves on the NSDA Coordinating Committee.

**Abigail Potter**
*Information Technology Specialist - Project Manager, Library of Congress*
Abigail Potter contributes to the National Digital Information Infrastructure and Preservation

Program at the Library of Congress. Currently, she is the NDSA Coordinating Committee facilitator. She also contributes to Viewshare.org, an open source web tool that provides enhanced access to digital collections, and to a series of content summits that explore new forms of born-digital content that libraries, archives and museums are collecting. Recently, Potter also served as the Communications Officer to the International Internet Preservation Consortium.

**Robin Ruggaber**
*Chief Technical Officer for the University of Virginia Library*
Ruggaber serves in a strategic or technical advisory capacity to Fedora, Blacklight, Hydra, and APTrust, and is responsible for the strategic, architectural and operational aspects of technology for the UVa Library. Ruggaber's career spans across industry, federal and state agencies with deepest expertise in directing the design and development of strategies and systems to solve complex problems and meet organizational objectives. She is drawn to work in digital stewardship due to the complex challenges facing the community and the opportunity to protect availability and access to intellectual and cultural knowledge.

**John Spencer**
*President and co-founder, BMS/Chace.*
John Spencer has widespread experience and visibility both in the music industry and in the fields of archival preservation and enterprise class information technology. Since 1978, he has been involved in many facets of high-technology professional audio and video, and was previously Vice President of Sales and Marketing for Otari Corp. Spencer is CEO of BMS/Chace, working with commercial companies to further the value of structured metadata collection for media companies and institutions worldwide. He is a member of the following professional associations: The Recording Academy Producers & Engineers Wing National Advisory Council, and Nashville chapter Advisory Council, Audio Engineering Society (AES) Studio Practices and Production Technical Committee, Association of Recorded Sound Collections (ARSC) Technical Committee, National Recording Preservation Board (NRPB) Digital Audio Preservation, and Standards Task Force. Spencer serves on the NSDA Coordinating Committee.

**Helen Tibbo**
*Alumni Distinguished Professor, School of Information and Library Science, University of North Carolina at Chapel Hill*
Professor Helen Tibbo teaches in the areas of archives and records management, digital curation, electronic retrieval, and reference. She is an SAA fellow, has served on SAA committees and boards for over 20 year, was the co-founder of the SAA Research Forum, and was SAA President, 2010-2011. She is currently the primary investigator for the IMLS-funded CRADLE project that is creating learning tools for archivists, librarians, and science researchers regarding data management. She was PI for DigCCurr I and II, which developed Digital Curation curriculums for master's level students (2006-2009), doctoral students and information professionals (2008-2013). Tibbo serves on the NSDA Coordinating Committee.

**Kate Wittenberg**
*Managing Director, Portico*

Before taking on the leadership of Portico, Kate Wittenberg was Project Director, Client and Partnership Development in Ithaka S+R, where she focused on building partnerships among scholars, publishers, libraries, technology providers, and scholarly societies with an interest in promoting the development and sustainability of digital scholarship and learning. Before joining Ithaka, she directed the Electronic Publishing Initiative at Columbia, a collaboration of the libraries, academic computing, and university press. Kate serves on the NDSA Coordinating Committee.