STEWARDING THE INVISIBLE

Setting the Stage for Institution-Wide Digital Preservation at the Smithsonian

Final Report
November 15, 2016

Submitted by

AVPreserve
253 36th St, Suite C302
Brooklyn, NY 11232
917-475-9630

Kara Van Malssen
kara@avpreserve.com

Chris Lacinak
chris@avpreserve.com

With contributions by Beth Delaney and Amy Rudersdorf, AVPreserve. Special thanks to Anne Van Camp, Director of the Smithsonian Institution Archives and chair of the pan-Smithsonian Digital Preservation Working Group, and all Working Group members.

TABLE OF CONTENTS

1 INTRODUCTION

Imagine being able to access all known information about an insect species – whether it was discovered 100 years or 100 days ago – with one touch of the screen. Picture a world in which you can not only see Smithsonian objects online but also hear them and watch them in motion. Or imagine learning that Smithsonian astrophysicists discovered a new, Earth-like planet orbiting a star five light-years away.

— Smithsonian Institution Strategic Plan, Fiscal Years 2010-2015[1]

The 2010-2015 Smithsonian Institution Strategic Plan laid out a grand vision for the future, one in which the vast trove of information collected and created by the Institution would be quickly and easily accessible to students, educators, enthusiasts, and professionals, enabling new knowledge to be generated through previously undiscovered interconnections between datasets, collection items, and other resources. This vision imagines the creation of a digital universe wherein such discoveries would be enabled, new research performed, new datasets generated and new collections acquired.

The Institution's goals for building this digital universe are ambitious. The complementary 2010-2015 Digitization Strategic Plan *Creating a Digital Smithsonian* laid out a framework for undertaking digitization "of our collections and research holdings along with the descriptive, interpretative information that accompanies them,"[2] in order to expand access to these resources in unparalleled ways. The creators of the plan realized, however, that it was not enough to simply digitize; for that digital universe to persist and grow long into the future, great care of digital resources would be required. The Digitization Plan recognized that:

> To avoid digitized materials becoming obsolete, we must digitize at the highest quality, migrate to the latest storage and formats, and maintain the links to the descriptive information that makes digital assets meaningful. The Smithsonian will take a life-cycle management approach to digitization based on carefully crafted standards and best practices that will ensure the highest fidelity and widest range of uses. We will keep a close eye on the changing technologies that the participatory web and new media world will surely bring. Equally important is establishing guidelines for disposing of data we no longer need to retain.[3]

[1] Smithsonian Institution. Inspiring generations through knowledge and discovery: Strategic plan, 2010-
[2] Smithsonian Institution. Creating a digital Smithsonian: Digitization strategic plan, 2010-2015. p. 2. https://www.si.edu/content/pdf/about/2010_SI_Digitization_Plan.pdf. Accessed September 26, 2016.
[3] Ibid., 3.

A great deal has been accomplished toward the goals of these two visionary documents. As of September 2016, the Smithsonian Institution Dashboard shows that there are:

- 1,954,315 museum objects and specimens represented by digital images
- 8,495 archival cubic feet represented by digital images
- 26,583 library volumes represented by digital images
- A combined total of 26,870,573 electronic records representing museum, library and archival holdings[4]

Aside from this progress on the digitization front, the born-digital output of research efforts is expanding, further contributing to the growth of the Smithsonian digital universe.

There is a tremendous amount of work to be done to cultivate a digital environment that ensures the digital resources that exist today, and those that are created tomorrow, remain available far into the future. Yet, it has been observed, as high up in the organization as the Secretary, that the lifecycle management and, more specifically, digital preservation component of the strategic vision has not yet been tackled in a coordinated manner, and that newly digitized as well as born-digital resources are at risk. In 2015, former Secretary of the Smithsonian, Wayne Clough, issued a memo stating, “We have made great progress in realizing the vision of a Digital Smithsonian. Along the way, we have worked at both the central and unit levels to address management challenges that have emerged in this dynamic arena. I would like to highlight one issue that will surely grow in size and complexity over the coming years: life-cycle management of digital data.” This memo introduced the Digital Preservation Working Group (DPWG) and charged it with assessing current practices and creating recommendations to improve the preservation of digital resources.

This report represents the outcome of the DPWG’s charge and presents the findings of the first Institution-wide assessment of current digital life-cycle management practices. The study was conducted by AVPreserve, a consulting and software development firm with deep expertise in digital preservation and enterprise data management, on behalf of, and under the guidance of, the DPWG. The authors are particularly grateful to Anne Van Camp, Director of the Smithsonian Institution Archives, for her leadership during this process.

Taking a broad view of Institutional digital resources, the study examines the current and future digital preservation needs and goals for collection items, research data, other forms of Institutional output, and the metadata that describes these. With input from a wide variety of stakeholders, it looks at digital preservation practices at individual, unit, and organization-wide levels, and identifies distinct challenges to the organizational alignment of these practices. Finally, recommendations are provided, outlining steps that the Smithsonian Institution can take to enact systemic preservation of its valuable resources and ensure that the vision for a digital future is realized in a managed and secure manner.

[4] As of September 12, 2016

The report is structured as follows:

- **Section 1 — Introduction**: Presents background information, scope, and methodology used for the study.
- **Section 2 — Findings: Interviews**: Reports and analyzes what we heard from interviewees regarding digital resource types and quantities, systems in use, current challenges, and ideas for improving the state of digital preservation.
- **Section 3 — Findings: Researcher Survey**: Summarizes the results of an online survey on research data holdings that was distributed to Smithsonian Researchers in August 2016.
- **Section 4 — Findings: Documentation Review**: Presents our analysis of several key documents that guide digital preservation today.
- **Section 5 — Summary of Findings**: Presents our conclusions based on the interviews, researcher survey, and document review.
- **Section 6 — Recommendations**: Provides our recommendations for initiating a pan-Institutional digital preservation program.

Short case studies, illustrating specific challenges or successes, can be found throughout the report.

Several appendices accompany the report:

- **Appendix A. Glossary:** Definitions for terms highlighted in orange throughout the report.
- **Appendix B. Stakeholders:** List of stakeholders interviewed, their units, and the date of the interview.
- **Appendix C. Interview Notes:** Summarized notes from each interview conducted.
- **Appendix D. Interview Analysis:** Analysis of interview responses.
- **Appendix E. Researcher Survey Data - Raw**
- **Appendix F. Researcher Survey Data - Summarized**

1.1 Background and Previous Work

At the Smithsonian Institution, digital preservation is not a new concept, yet it is also not a formalized program at the enterprise level…Unlike the SI Directive for digitization (SD 610), there is no explicit SI Directive for digital preservation at this point…There is an opportunity in the coming years for a unified approach to digital preservation at SI, much like what has been done with digitization through the DPO.

— AVPreserve, Smithsonian Institution DAM Digital Preservation Assessment (2015)

This study follows a digital preservation assessment of the Smithsonian Enterprise DAMS, conducted by AVPreserve in 2015, which was based on the international standard Audit and Certification of Trustworthy Repositories (ISO 16363). The preservation assessment was conducted in order to evaluate the trustworthiness of what had become a *de facto* preservation system for the Institution. While the original intent of the DAMS was to provide access to digital resources, user behaviors and expectations placed the burden of preservation on the system, bolstered by the lack of any alternative system designated as a preservation environment. Over time, the DAMS has became a repository of unique data, representing assets of the Institution for which there are long-term access requirements and expectations. A third-party standards-based assessment was determined to be the best way to characterize the maturity of the DAMS as a preservation system, and could identify the necessary improvements to bring it to full compliance with industry standards.

The assessment found that the DAMS functions as a mature and robust digital preservation system, with particular strength in the Digital Object Management procedures, and Infrastructure and Security Risk Management metrics defined by ISO 16363. The majority of its shortcomings were found in the Organizational Infrastructure section of the standard, which has metrics to evaluate the repository's operational guidance, maintenance, and sustainability. It includes fundamental elements of the repository's mission, strategic plan and policy framework, staffing and budgetary considerations, accountability measures, and documented agreements between the repository and its depositors[5].

The assessment concluded that these deficient areas were largely the result of an unclear policy and procedural framework *at the Institutional level*. It concluded that without clarity on these issues from Smithsonian administration, the DAMS would continue to be challenged to resolve these gaps. The report highlighted this issue as central, noting, "Despite the considerable investment and expressed support of the system from the OCIO, the lack of an official SI-wide preservation mandate, a clear definition of scope, and a clear expression of the DAMS service model and associated strategies continues to cast doubt on the DAMS's role as a repository for long-term preservation."[6]

This study picks up the issue from this point, looking across the Institution to understand where the gaps in digital preservation are, why they exist, and what can be done to resolve these and move forward.

1.2 Goals and Objectives

The goal of this study, as charged by the Digital Preservation Working Group (DPWG) is to identify gaps in digital preservation responsibility, coordination, and policy, and to identify possible solutions that will enable the Smithsonian to move toward systematic preservation of all digital resources of enduring value. Uncovering these gaps and identifying short and long-term

[5] DAMS digital preservation assessment final report. September 1, 2015. p. 25.
[6] Ibid., 40.

solutions is a first step toward the broader goal of making digital preservation a seamless and organic underlying function of the Institution.

The study leverages and relates the voices of stakeholders from across the Institution in order to identify shared challenges and offer solutions that reflect the ideas of those deeply committed to the Smithsonian's coordinated preservation vision. This is one in which every stakeholder, from content creators to collection managers to repositories, understand the overarching goals and their responsibility in reaching those goals. A set of key questions guide the analysis toward this end.

Goal

To identify gaps in digital preservation responsibility, coordination, technology, and policy, and to identify possible solutions that will enable the Smithsonian to move toward systematic preservation of all digital resources of enduring value.

Key Questions

- What digital resources can be found across the Institution?
- How are they being stewarded today?
- What factors are currently inhibiting systematic Institution-wide digital preservation?
- How can the organization align and work cooperatively toward fulfilling the digital preservation charge?

These questions support the DPWG's charge by seeking to identify potential resources of value, both found in museum, library, and archive collections, as well as within research departments across all sectors of the Institution, and understanding what steps can be taken to ensure their preservation. Specifically, the report seeks to further the Smithsonian's vision toward achieving the goal of coordinated preservation, by pursuing several overarching objectives:

Recognize the role of digital preservation within strategic vision and initiatives. Situating the task of improving digital preservation within the current strategic plan provides a starting point for conversation by anchoring this study in the Institution's vision and tying the effort toward current digital activities being undertaken at the museum, such as large-scale digitization of collections.

Establish a common vocabulary and reference points. This includes definition of "digital preservation" and proposed complementary vocabulary to establish a common framework for discussion and action. Definitions must be tangible and understood by a diverse group of

stakeholders throughout the Institution so they are widely adopted and used consistently. This will foster productive conversations between and amongst stakeholders, including collections and research communities, Senior Leadership, and supporting technologists. Definitions for digital resource types should be agreed upon in order to provide clarity around currently hazy designations.

Understand digital preservation challenges and goals from stakeholders' perspectives. Looking at the Institution's preservation needs, goals, and challenges from the perspective of a large and diverse group of stakeholders presents the opportunity to understand and empathize with many unique vantage points, as well as to assemble these into a high-level 360 degree picture of the landscape.

Situate the Smithsonian within standard models of digital preservation maturity and identify gaps. Digital preservation on the surface appears to be a fairly simple set of processes. But behind any well-functioning preservation pipeline is a complex set of decisions, actions taken by numerous diverse stakeholders, and a robust technical foundation. Examining the current state of digital preservation practice at the Smithsonian against best practices that account for these interrelated factors will make underlying issues visible.

Provide recommendations to address gaps and risks, and obtain Institution-wide goals. Coordinated and concerted effort will be required in order to accomplish the necessary shifts across the Institution and achieve long-term goals. A set of recommendations will provide a foundation for development of specific action items.

1.3 Scope

This study is very broad in many respects, yet specific in others. Here we describe the report scope using several parameters: functional, content, and organizational. We also use this opportunity to further explore several concepts that will feature prominently in the report. They establish a baseline set of definitions that will be used throughout, and ideally in the next phases of this initiative.

Note that all terms highlighted in orange can also be found in **Appendix A. Glossary.**

1.3.1 Functional Scope

This study evaluates how mature the Smithsonian's digital preservation capabilities are, and determines what areas of improvement can be made to move toward a coordinated, Institution-wide program. Digital preservation, "refers to the series of managed activities necessary to ensure continued access to digital materials for as long as necessary."[7] Digital preservation actions are taken over time throughout changing organizational and technological environments

[7] Digital Preservation Coalition. Introduction — Definitions and Concepts. http://handbook.dpconline.org/glossary#D. Accessed September 26, 2016.

to enable ongoing accessibility and usability of digital resources that contain valuable content. Digital preservation serves the function of risk management for digital data to ensure that they can still be found and used, despite shifts in operating systems, software, file formats, and hardware that will inevitably occur. Digital preservation actions include unique identification of digital assets; establishment and validation of asset integrity and fixity; secure storage, backup, and disaster recovery; ongoing monitoring and threat mitigation; security management; and delivery of files to appropriate users and/or use environments. These actions are informed by policies and enforced by people and technologies.

The concept of continued access is a critical component of this definition. If preservation actions are performed in a vacuum, without ties to access, preservation is not being achieved. Therefore, while this report is not evaluating, for example, how well digital resources are being made available today, or how efficient the digitization process is, these topics come under consideration to the extent that they have an impact on digital preservation activities, and that digital preservation activities have an impact on them.

The term life-cycle management is sometimes used to imply digital preservation; in fact this concept encompasses a broader set of activities including decision-making about what to preserve, and disposition once retention periods have ended. Preservation is one core activity in life-cycle management.

Digital asset management is also complementary to digital preservation, and is defined as the, "management tasks and decisions surrounding the ingestion, annotation, cataloguing, storage, retrieval and distribution of digital assets."[8] Such activities and systems typically support workflow management, collaboration, and access goals, aimed at serving immediate user needs for access to file-based assets. However, digital asset management alone does not ensure preservation. Digital preservation activities center on maintaining the option to serve current *and* future users — by ensuring fixity, persistence of files, and authenticity of content. Given the shared goal of access, however, It is not uncommon to find digital asset management systems that have been expanded to perform preservation functions, such as in the case of the Smithsonian Institution DAMS.

Any technical environment composed of a storage layer, database for metadata management, and access interface, and has dedicated staff that provides access, management, and/or preservation services will be referred to in this report as a repository. Repositories are distinct from purely storage environments, which only provide access through the file system, and do not provide features such as search, browse, and description.

[8] Digital Assets Management. Wikipedia. http://en.wikipedia.org/wiki/Digital_asset_management. Accessed September 26, 2016.

1.3.2 Content Scope

This study assumes that all digital resources of the Institution have *potential* long-term value, and are therefore within scope. Digital resources is a broad concept, intended to be used as a catch-all for the various types of digital information found throughout the Institution.

Digital resources are composed of digital files, of which there may be multiple that constitute a given resource (e.g., multiple images of a vase from different angles, multiple recordings of bird sounds created during field research). The content of digital files are considered to be a type of data, specifically a type of unstructured data.[9] Digital files are a distinct data type in that they are composed of bytes, including a beginning and ending byte, and have an address that is known to an operating system. Structured data, those with a data model, are what we think of as the values in a database. Therefore, an image file and numeric strings in a database can equally be considered data. We use the term data when appropriate in this report to largely imply the same meaning as digital resources.

Digital assets are considered to be digital resources that are identified to have enduring value to an organization, due to their potential for ongoing use in meeting strategic objectives and re-use in the creation of other digital resources. The designation of "asset" is highly dependent on policy and selection criteria. Just because everything could potentially be in scope for preservation, doesn't mean it all realistically should be, and therefore policies and selection criteria are important to understanding the difference between digital resources and digital assets. The value designation of assets is evidenced by the ability for users to identify, find, and utilize them effectively, in large part due to the accompaniment of metadata, and availability on networked systems.

Although determining what is and is not an asset of the Smithsonian is outside the scope of this study, establishing a clear terminological framework in support of discussion and action that helps navigate this path is an objective.

Digital Collection Items

The term collection item is based on the concept from Smithsonian Directive 600 — Collections Management (SD 600), which states:

> Smithsonian holdings include museum, archive, and library collections. Collections may be categorized by legal and curatorial status and the intended use of the

[9] Wikipedia defines unstructured data as, "information that either does not have a pre-defined data model or is not organized in a pre-defined manner." The article further elaborates that, "Examples of "unstructured data" may include books, journals, documents, metadata, health records, audio, video, analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document. While the main content being conveyed does not have a defined structure, it generally comes packaged in objects (e.g., in files or documents, ...) that themselves have structure and are thus a mix of structured and unstructured data, but collectively this is still referred to as "unstructured data." https://en.wikipedia.org/wiki/Unstructured_data. Accessed September 26, 2016.

> collections. Collections include items (referred to here as 'collection items') acquired for accessioned, non- accessioned, supplementary, study, or research collections, provided the items are acquired, preserved, and maintained for public exhibition, education, or research.[10]

Digital collection items include both digital surrogates of physical collection items that are created through digitization, as well as born-digital collection items, or those for also known as Primary Digital Collection Objects (PDCO).

For the purposes of this study, we consider collection items to be those that are accessioned into a museum, library, or archive according to the collection policy of those units. Un-accessioned and ad hoc collections will be considered Institutional output, as described below.

Research Data

This category includes, "Any digital data that is collected, observed, or created for purposes of analysis to produce original research results."[11] This data can be further broken down into several categories, an example of which can be seen in Figure 1.

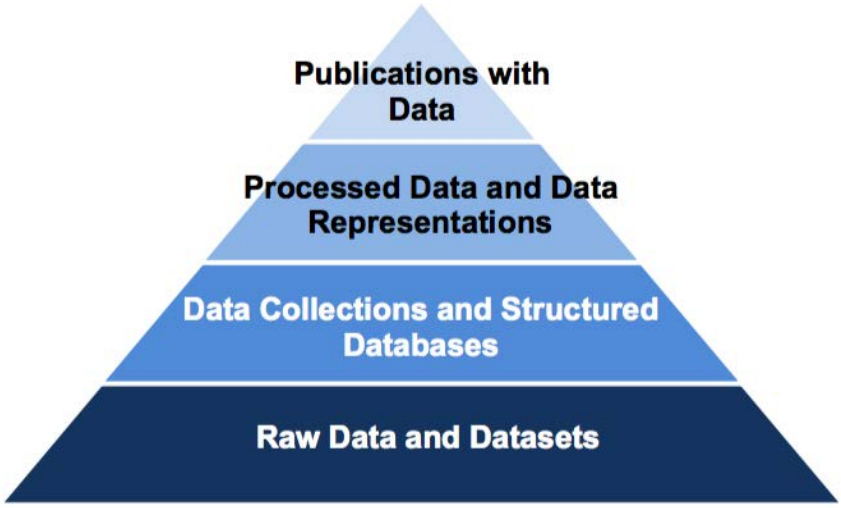


Figure 1. MPS Levels of Data.[12]

For the purposes of this study, we use a similar categorization to define "research data" at the Smithsonian:

[10] Smithsonian Institution. Smithsonian Directive 600: Collections management, October 26, 2001. p. 8. https://www.si.edu/content/pdf/about/sd/SD 600andAppendix.pdf. Accessed September 26, 2016.
[11] Boston University Libraries, Research Data Management. What is "research data"? http://www.bu.edu/datamanagement/background/whatisdata/. Accessed September 26, 2016.
[12] Hanish, Robert, et. al. MPS Open Data workshop series draft report. MPS Open Data. https://mpsopendata.crc.nd.edu/images/Reports/MPS_ReportDraft_v4.pdf. Accessed September 26, 2016.

- Raw / Primary data: Data generated through observations, instruments, and/or experiments, either as originally collected or after being checked, calibrated, and/or organized (e.g., survey responses, sensor data, neurological images). This is the bottom two layers of the pyramid in Figure 1.

- Analyzed / Derived data: Refined data derived from either the researcher's own or from third party Raw / Primary research data and interpreted by the researcher through some form of manipulation, transformation, or abstraction (e.g., statistical analysis or modeling). This is represented by the "processed data and data representations" layer in Figure 1.

- Publication data: Reference or canonical data that is a subset of Analyzed / Derived data and is likely peer reviewed, published, and/or curated. Publication data is potentially a synthesis of several researcher's ideas and datasets. This is the very top of the pyramid.

Associated Information

Finally, we must carefully consider the terms that will help our audience both find objects and add meaning. Ask for soda and depending where you are in America, you might get sodium bicarbonate, seltzer, or a soft drink. A fourth-grader completing a homework assignment on the night sky might key in "star," whereas a post-doctoral fellow or scholar would use more specific and sophisticated terms. Judaica scholars might search the collections using Hebrew or English. If the science — and expense — of digitization rests with the technology, then much of the art relates to vocabulary. Language changes constantly, and as access broadens, the public will offer infinitely more descriptions per item.

— Smithsonian Institution Digitization Strategic Plan, Fiscal Years 2010-2015[13]

SD 600 defines collections information as follows:

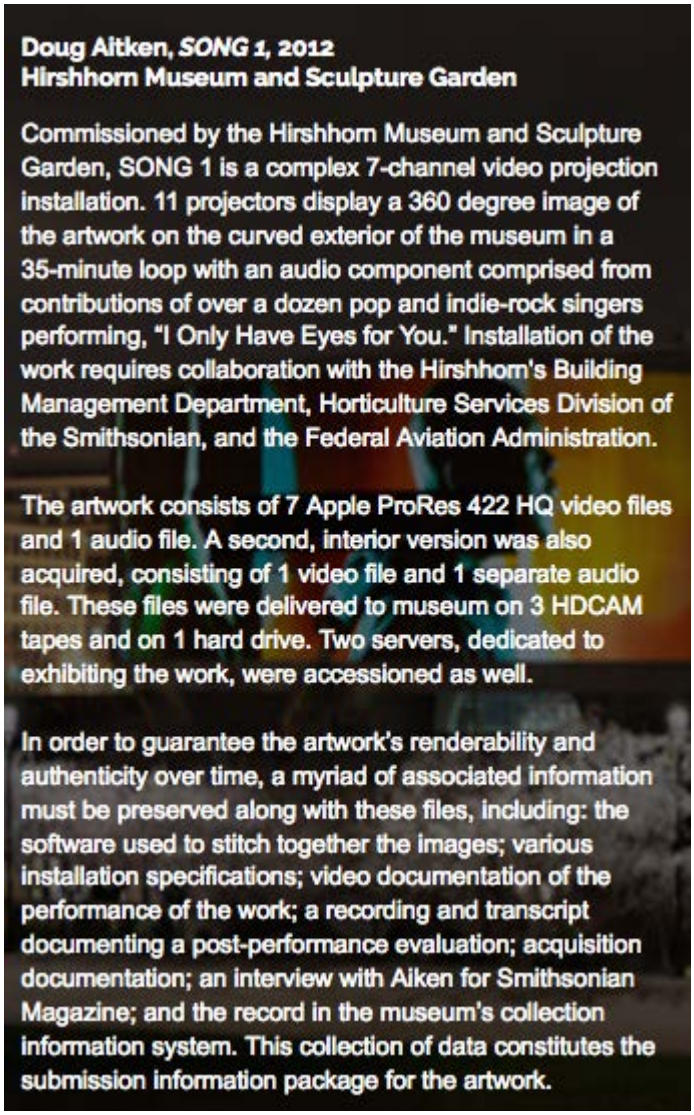
> The primary purpose of collections information is to provide access to Smithsonian collections, research findings, and the stories they can tell. To support this goal, the Smithsonian has a responsibility to acquire, develop, and maintain collections information systems that enhance access to and accountability for its collections and research findings and to ensure long-term preservation of the resultant information in manual and electronic formats.[14]

[13] Smithsonian Institution. Inspiring generations through knowledge and discovery: Strategic plan, 2010-2015. p. 6. https://siarchives.si.edu/sites/default/files/pdfs/SI_Strategic_Plan_2010-2015.pdf. Accessed September 26, 2016.
[14] Smithsonian Institution. Smithsonian Directive 600: Collections management, October 26, 2001. p. 16-17. https://www.si.edu/content/pdf/about/sd/SD 600andAppendix.pdf. Accessed September 26, 2016.

SD 600 also uses the term associated information to have a similar meaning, which is the preferred term that will be used in this report to refer to any data that surrounds collection objects or research data, for the purpose of supporting their accessibility and stewardship. This includes metadata stored in Collections Information Systems (CISs), repository databases, spreadsheets, or other data storage formats. It also includes documentation that supports the above goals, such as installation instructions for an artwork, or even dedicated software required to display the work. Even digital resources themselves may be considered a form of metadata: digital images of a collections item should be considered collections documentation. Associated information includes metadata that surrounds research datasets, such as the location where a set of primary field data was collected, who collected it, and on what date and time.

For digital resources to remain accessible over the long-term, they must be findable, understandable, and usable to the community for whom they are being preserved. Therefore, the associated information surrounding a digital resource must also be preserved. This idea of preserving all the information required to ensure long-term access to a digital resource is well articulated by the standard OAIS Reference Model,[15] which refers to these datasets as information packages. These are logical packages, which exist in three possible different versions as they are transformed throughout the preservation life cycle:

- Submission Information Package (SIP): The SIP is the information package as acquired by the repository from the content producer (which could be a content creator, digitization service, or any other entity responsible for creating digital resources). It contains all digital files that make up a content item, as well as provided metadata.
- Archival Information Package



[15] ISO 14721:2012. Space data and information transfer systems — Open archival information system (OAIS) Reference model. https://public.ccsds.org/pubs/650x0m2.pdf. Accessed September 28, 2016.

(AIP): In order to manage the content over time, a preservation environment will add additional technical, structural or preservation metadata to that submitted by the producer to create a complete set of preservation information. The preservation environment in which the information package is managed continues to add this information over time so that its activities can be traced, and the provenance and authenticity of content can be guaranteed. This collection of data makes up the AIP.

- Dissemination Information Package (DIP): DIPs are created from AIPs to serve access needs. These are typically portions of an AIP, which vary in detail depending on the usage required. They are created either on systematically or on demand.

This study takes into consideration both primary digital resources as well as any important associated information that would contribute to the ability to locate, utilize, understand, and manage those resources. The sum of this information may be referred to throughout the report as an information package, or one of the three variants.

Institutional Output

There are large volumes of digital resources at the Smithsonian that do not fall under the above categories that may also have long-term value. For lack of a better term, these resources will be referred to as institutional output. This category could include a broad range of content, such as human resource records, event videos and photographs, architectural models, marketing photographs, and more. The gray areas between institutional output, collection items, research data, and associated information will be explored later in this report.

1.3.3 Organizational Scope

This study and its findings are applicable to the entire Smithsonian Institution, including the 19 museums, 9 research centers, the National Zoo, as well as the administrative offices and supporting services, staff, and facilities.

1.4 Methodology

This report is the product of several months of interviews with stakeholders across the Smithsonian; research into existing strategy, policy, and procedure; and an online survey of the Institution's researchers. Resulting qualitative and quantitative datasets were then analyzed. Further detail about each approach is provided below. All collected information, including a complete list of interviewees, interview notes, survey results, and analysis, are available in the appendices.

1.4.1 Interviews

Between a four-day onsite visit on April 25-28, 2016, and additional phone interviews from May - July of 2016, AVPreserve conducted interviews with 16 stakeholder groups. The stakeholders

provide a sampling from across the Institution representing three different perspectives on digital preservation: Collecting Units, Repositories, and Content Producers.

The interview approach was not intended to be exhaustive, but to provide a snapshot of the current state and help illuminate major gap areas. Findings from the interviews are summarized in Section 2. A full list of stakeholders interviewed is provided in **Appendix B. Stakeholders.** Edited notes from the interviews are available in **Appendix C. Interview Notes**.

Collecting Units: This category includes representatives from individual museum, library, or archives within Institutional units, who currently play some role in the stewardship of digital resources, and are invested in their longevity. Eight collecting units were interviewed, as well as the National Collections program, as a representative of collecting units, for a total of nine groups in this category:

- Arthur M. Sackler and Freer Gallery of Art (Freer | Sackler)
- National Museum of American History (NMAH)
- National Museum of the American Indian (NMAI)
- National Museum of Natural History (NMNH)
- Hirshhorn Museum and Sculpture Garden
- Smithsonian Center for Folklife and Cultural Heritage (CFCH)
- Smithsonian Institution Archives (SIA)
- Smithsonian Institution Libraries (SIL)
- National Collections Program (NCP)*

Repositories: This category includes representatives who either provide technological services for the management of digital resources directly to cross-Institutional constituents, or who take responsibility for cross-Institutional digital resources and manage their technological oversight. Representatives from four repositories were interviewed:

- Smithsonian Institution Enterprise Digital Asset Management System (DAMS), OCIO
- SIdora, OCIO
- DSpace Digital Repository, Smithsonian Institution Libraries
- Smithsonian Digital Archives, Smithsonian Institution Archives

Content Producers: These are stakeholders whose primary role with regard to the stewardship of digital resources is as their creator, and therefore, are invested in ensuring their longevity. Five individuals and groups were interviewed, along with the coordinator of the Pan-Smithsonian Cryo Initiative, serving as a representative of researchers who deal with frozen specimens.

- Nicholas Pyenson, Curator of Fossil Marine Mammals, NMNH
- Robert Costello, National Outreach Program Manager, NMNH
- 3D Imaging Team, Digitization Program Office, OCIO
- Smithsonian Facilities Technology Advisory Council

- Pan-Smithsonian Cryo Initiative

Note that all Smithsonian staff should be considered content producers as they all create content that can be considered institutional output. Collecting units are sometimes referenced in this report as content creators, particularly due to their digitization activities.

1.4.2 Researcher Survey

A survey was conducted as part of this study in order to gather more comprehensive data on the types and quantities of research data actively being produced today, and how data is currently being managed. The survey results are particularly important because there is no current data available about how much output the Institution's research staff and fellows are generating, how much they anticipate creating in the next few years, how it is being stored, and how long it should be kept. More detail on the survey can be found in Section 3.

The survey was distributed to unit heads and research department chairs, who asked their constituents to participate. One hundred respondents completed the survey, out of 149 who started it or only completed part of it.

1.4.3 Documentation Review

In addition to gathering information from stakeholders, an important component of this study has been the review of existing policy, procedure, and strategy documentation, as well as previous analysis relevant to the current scope. The following documents, were closely examined (in Section 4) as they proved particularly relevant to the study:

- Digital Preservation Working Group Charge from Secretary Clough
- Smithsonian Institution Strategic Plan, Fiscal Years 2010-2015[16]
- Smithsonian Institution Digitization Strategic Plan, Fiscal Years 2010-2015[17]
- Smithsonian Directive 600 — Collection Management (2001)[18]
- Smithsonian Directive 609 — Digital Asset Access and Use (2011)[19]
- Smithsonian Directive 610 — Digitization and Digital Asset Management Policy (2011)[20]
- Smithsonian Directive 503 — Management of Archives and Special Collections at the Smithsonian Institution (2010), and related Smithsonian Directives 501 (1985) and 505 (1985)
- Several sample Digital Asset Management Plans (DAMPS) (2013)
- Concern at the Core: Managing Smithsonian Collections (2005)[21]

[16] Our understanding is that the 2010-2015 Strategic Plans have been extended to 2017. See Smithsonian Institution. Inspiring generations through knowledge and discovery: Strategic plan, 2010-2015. p. 2. https://siarchives.si.edu/sites/default/files/pdfs/SI_Strategic_Plan_2010-2015.pdf. Accessed September 26, 2016.
[17] https://www.si.edu/content/pdf/about/2010_SI_Digitization_Plan.pdf. Accessed September 26, 2016.
[18] https://www.si.edu/content/pdf/about/sd/SD 600andAppendix.pdf. Accessed September 26, 2016
[19] https://www.si.edu/content/pdf/about/sd/SD 600andAppendix.pdf. Accessed September 26, 2016
[20] https://www.si.edu/content/pdf/about/sd/SD 610.pdf. Accessed September 26, 2016.

- Sharing Smithsonian Digital Scientific Research from Biology (2011)[22]
- Smithsonian Plan for Increased Public Access to Results of Federally Funded Research[23] (2015)

1.4.4 Analysis

The interview results, survey results, and documentation review, were consolidated and analyzed (Section 5, "Summary of Findings") to identify gaps between current practice and future goals, and their associated risks and opportunities. Based on this analysis, we created a set of recommended strategies (Section 6) to advance the Smithsonian's goals towards a sustainable approach to Institution-wide digital preservation of its digital resources.

1.5 Evaluation Framework

> *Ensuring that valuable digital assets will be available for future use is not simply a matter of finding sufficient funds. It is about mobilizing resources — human, technical, and financial — across a spectrum of stakeholders diffuse over both space and time.*
> *— Blue Ribbon Task Force on Sustainable Digital Preservation and Access*[24]

Cornell University and MIT's long-standing Digital Preservation Management tutorial[25] proposes that the foundation of a viable digital preservation program is like a three-legged stool, with one leg each for organizational infrastructure, technological infrastructure, and resources framework, defined as follows:

- **Organizational Infrastructure** includes the policies, procedures, practices, people — the elements that any programmatic area needs to thrive, but specialized to address digital preservation requirements.

- **Technological Infrastructure** consists of the requisite equipment, software, hardware, a secure environment, and skills to establish and maintain the digital preservation program. It anticipates and responds wisely to changing technology.

---

[21] https://www.si.edu/opanda/policy
[22] http://www.si.edu/content/opanda/docs/rpts2011/11.03.datasharing.final.pdf
[23] https://www.si.edu/content/pdf/about/SmithsonianPublicAccessPlan.pdf
[24] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Final report of the Blue Ribbon Task Force on sustainable digital preservation and access, February 2010. p.1. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf. Accessed September 28, 2016.
[25] Digital preservation management: Implementing short-term strategies for long-term problems, Cornell University and MIT. http://www.dpworkshop.org/. Accessed September 26, 2016.

- **Resources Framework** addresses the requisite startup, ongoing, and contingency funding to enable and sustain the digital preservation program.

The stool analogy provides an excellent framework for evaluating a digital preservation program: take any one of these elements away and the program can't be sustained; cut one short and it becomes unstable.

The primary authors of the Digital Preservation Management workshop, Anne R. Kenney and Nancy McGovern, leverage this model in their seminal paper, "The Five Organizational Stages of Digital Preservation," which identifies five stages of organizational response to digital preservation that emerge as a result of increased experience, with key indicators at each stage for the three components of the stool model. The levels are:

1. **Acknowledge**: Understand that digital preservation is a local concern;
2. **Act**: Initiate digital preservation projects;
3. **Consolidate**: Segue from projects to programs;
4. **Institutionalize**: Incorporate the larger environment; and
5. **Externalize**: Embrace inter-Institutional collaboration and dependency.[26]

Within this study we use the 5 stage model as a starting point for looking at the readiness of the Smithsonian's digital preservation initiatives, assessing the organizational, technical, and resources foundation that are required for the Institution to launch a large-scale digital preservation program that is inclusive of collections, research data, Institutional output, and associated information. It will be referenced in Section 5 of this document.

[26] Kenney, Anne R. and Nancy Y. McGovern, "The five organizational stages of digital preservation," *Digital libraries: A vision for the 21st century,* 2003. Ann Arbor, MI: Michigan Publishing, University of Michigan Library. http://quod.lib.umich.edu/s/spobooks/bbv9812.0001.001/1:11/--digital-libraries-a-vision-for-the-21st-century?rgn=div1;view=fulltext. Accessed September 26, 2016.

2 FINDINGS: INTERVIEWS

A critical goal of this study is to understand the current state of digital preservation and goals for the future from the stakeholders' perspectives. In this section, we report what we *heard*. Where further research into a topic was required to gain clarity, those findings are reported here as well. Edited notes from the interviews can be found in **Appendix D: Interview Notes.**

The 16 stakeholder groups interviewed represented a variety of perspectives on the issue of Institution-wide preservation. Although discussion topics varied depending on the role of each stakeholder group interviewed, there were several common and consistent themes that were discussed with each group.

Interview Topics & Questions

1. **Resource types and quantities**: What types of data is your group (or your constituents) responsible for creating and/or stewarding, and how much is there?
2. **Systems:** What technical systems are being used for the management of data? What is the role of each system?
3. **Current obstacles or issues:** What are the biggest challenges today that inhibit comprehensive digital preservation?
4. **Ideas or opportunities for improvement:** What changes could be made to improve preservation of digital resources at the Smithsonian?

Findings on these common topics as relayed through interviews are summarized below.

2.1 Resource Types and Quantities

Key questions: What types of data is your group (or your constituents) responsible for creating and/or stewarding, and how much is there?

2.1.1 Resource Types

While an exhaustive list of the types of digital resources created or stewarded by each group could not be obtained during each hour-long interview, interviewees did discuss in general the digital resources they have some responsibility for, and put particular emphasis on those types they either identified as essential to preserve or that they have concerns about.

The table below represents a summary of the digital resource types that stakeholders perceive as relevant to a preservation context. These general types are categorized according to whether they were characterized in the discussion as a type of collections item, an example of research data, or of institutional output. It is important to note that while this list captures a number of resource types, *this is not an exhaustive list of the digital resource types found at the Smithsonian.* Nor is it complete for each category (e.g., collection item, research data, Institutional output) — there may very well be 2D physical object surrogates that are considered research data. The list is limited only to what was communicated during interviews.

Note that, to the greatest extent possible, an "x" in one of the columns below is based on how interviewees interpreted unit-level policy. In other words, while event recordings may be considered collection items according to one unit's policy (e.g., Folklife Festival recordings collected by the Center for Folklife and Cultural Heritage) they are considered institutional output by another, therefore, both categories are checked.

Resource Type	Collection Item	Research Data	Institutional Output
2D physical object surrogates	x		
3D physical object surrogates	x	x	
Scanned photographic material	x		x
Born digital imagery	x	x	x
Digitized film, video and audio	x		x
Born digital video and audio	x	x	x
Scanned books and manuscripts	x	x	
Born digital artworks / design	x		
Digitized time based media artworks	x		
Email	x	x	x
Born digital documents	x	x	x
Design files			x
Websites	x		x
Social media	x		
CAD	x		x
GIS		x	x
Software/scripts	x	x	
Text (not documents)		x	
Metadata/records	x	x	x

Table 1. Resource types discussed by interviewees, categorized according to whether they were characterized by interviewees as collections items, research data, or Institutional output.

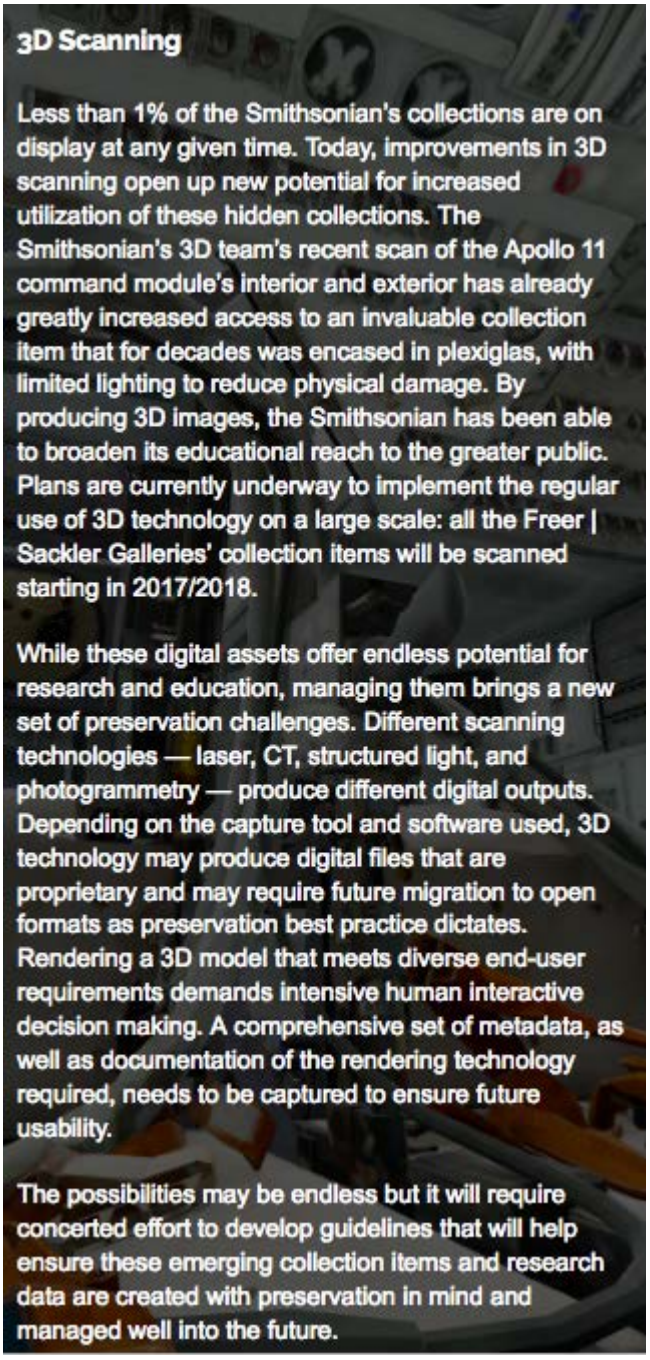


Table 1 demonstrates that there is a wide variety of digital resources across the Institution, which are created for a range of purposes, that interviewees feel are of value for long-term preservation. Furthermore, there is nothing inherent about a particular resource type that puts it in one category or another; digital images, email, and documents were discussed as important parts of collections, research, and Institutional output. Complicating this picture, and what is not represented here, is that the lines between these three categories are fluid throughout a resource's lifecycle. For example, email created by executive offices is institutional output that become collection items once acquired by the Smithsonian Institution Archives. Likewise, digital surrogates of the National Museum of Natural History's collections become the subject of study and contribute to the generation of research data.

The gray area between the different resource categories can result in a great deal of confusion for stakeholders. One interviewee asked: should the research data that is based off of an NMNH collection be accessioned and related to that collection as well? This confusion contributes to the challenges surrounding digital preservation because, in the digital world, it becomes difficult to draw distinct lines between resource types. In theory, this should not be an issue, after all, these are all just digital files. But in practice it is: collections are afforded funding and other resources that support their preservation that are currently unavailable to data considered the product of research or institutional output.

It should also be noted that nearly all interviewees raised the topic of metadata in the preservation context. Many they consider metadata of equal, if not at times greater, importance than the data they reference, and therefore must be a target of preservation themselves. They feel that currently metadata isn't viewed and prioritized as a type of asset, which places

longevity of the Institution's digital resources at risk of not being findable or understandable in the future. Metadata is therefore included in the table above in all three categories.

2.1.2 Quantities

In addition to types, quantities of data were also discussed. Interviewees addressed this in a number of ways, by reporting metrics such as:

- Number of collection items, which may consist of one or more digital files (e.g., for multiple pages in a book, or multiple views of an object)
- Number of files
- Volume of storage utilized

Given the small sample of stakeholders interviewed, and the diverse ways that quantities were discussed, we have approached this question from a number of perspectives. For collections, we received digitization and prioritization numbers from the National Collections Program, which tracks collections digitization efforts (see table 2, below). For a mixed perspective on collections, research data, and other institutional output, we looked at total repository volumes (see Section 2.2.2 Connections Systems in Use, below). For research data specifically, the results of the Research Data Survey (see Section 3 Findings: Researcher Survey, below) were used to estimate quantity. In this section, we focus on the quantities of digitized and born-digital collections, as those other numbers will be discussed further in the report. An estimate of current total volumes will be presented in Section 4 Summary of Findings.

At the end of fiscal year 2015, the following totals were reported to the National Collections Program office:

Collection Type	Digitized	Prioritized for Digitization
Museum	2,452,288 objects and specimens	13,099,799 objects and specimens
Archive	25,962 cubic feet	80,044 cubic feet[27]
Library[28]	28,133 volumes	696,402 volumes

Table 2. Smithsonian digitization and prioritization at the end of FY 2015. Provided by the National Collections Program.

These numbers only tell part of the story of the digital collection items. They do not tell us what volume of storage these digital assets require. And they do not reflect the accessioning of born-digital collection items, otherwise known as Primary Digital Collection Objects (PDCO) by museums, libraries, and archives. The National Collections Program reported that currently

[27] Note that the prioritized cubic feet of SIA digitization exceeds by 5 times the total claimed cubic feet of all archival groups at the Smithsonian (153,121 cubic feet).
[28] Totals limited to library books only, and does not include collections such as trade publications or special collections.

there is only preliminary data on PDCO holdings, which to date has only been collected from a handful of archives. PDCO holdings of museums and libraries have not yet been reported, meaning, for example, there are no current totals of born-digital software and time-based media art works held by the various fine art museums.[29]

The data in Table 3 have been provided by the National Collections Program, and reflect the total PDCO volumes by unit archive.

Unit	GB Total (Required)	# Objects (Optional)	# Files (Optional)	# Media (Optional)	# Hrs/Mins (Optional)
AAA	18,273	2,261	10,525	2	7,616
CFCH	29,829	0	262,831	0	0
NASM	115,226	7,838	0	0	80
NMAfA	215	0	0	0	0
NMNH	156	0	51,387	0	0
SG	515	0	16,468	0	0
SIA	8,439	0	0	0	0
Grand Total	172,652	10,099	341,211	2	7,696

Table 3. Seven Archival Collections voluntarily reported, 12 collection subsets with PDCO material. June 27, 2016 Analysis based on CDRS data FY 2015 V1.0 FINAL. Provided by Bill Tompkins, National Collections Program.

As can be seen in the table above, 172 TB of PDCO data has already been accessioned by unit archives. That number will likely grow significantly once museum and library collections are factored in, especially given that many born-digital artworks are video or film-based, which can mean large file sizes and complex information packages.

Given so many unknowns (and at times incompatible metrics for measuring quantity) it is difficult to make more than generalized conclusions about current digital collection holdings and growth for PDCO items and digital surrogates. One thing that is certain, and was confirmed by interviewees, is that the numbers of digitized and born-digital collection items will be increasing dramatically over the coming years. Interviewees expressed concern about this growth for three reasons. The first is that managed, enterprise storage capacity will need to keep pace with growth (see Section 2.2.1 Repositories, below). Most people we spoke to feel this is outside of

[29] As far as we can gather based on the interviews. The Time Based Media Art Working Group appears to have conducted a survey of these works in 2011-2012, but the numbers today are undoubtedly much higher than they were 5 years ago. One museum reported that the rate of acquisition of born-digital works has increased “exponentially” in the past few years.

their control, and they are already struggling to find ample storage. The second is that all of these digital resources must be managed if they are to persist over the long-term, and they are already finding the management of current volumes a challenge (see Section 2.3 Current Obstacles or Issues, below). The third is that in order for these digital collection items to be made accessible immediately and in the future, extensive accompanying metadata must be generated, which staff find a challenging mandate to fulfill given current staffing constraints (see Section 2.3 Current Obstacles or Issues, below).

The accumulation of new digital backlogs (which is likely to occur if staffing resources are not available to process these items) presents tremendous risk to the longevity of content, as these items are very likely to become "orphaned" over time. The largely unknown quantities of collection items further complicates the picture because it leaves the door open to complacency, overwhelms, or causes fear, which in turn causes people to turn away from the issue and hope it will be taken care of by someone else.

2.2 Systems

Key questions: What technical systems are being used for the management of data? What is the role of each system?

A variety of technical systems are employed across the Smithsonian for the management of digital resources over their life cycle. Interviews were conducted with representatives of several key systems in order to understand the current function and capacity of each. We also asked stakeholders to describe what systems they use for the management of digital resources to complement the vantage point of the system administrators with that of their end users. Each of the environments discussed play a role in the short- and/or long-term retention of one of the following digital resource categories under assessment in this study: digital collection items, research data, institutional output, or associated information. This includes the digital asset and data repositories, storage environments, and collection information systems (CISs).

In this section, we look at these these systems from two perspectives. The first is to summarize the existing digital repositories that specifically function as services to cross-SI stakeholders, or hold digital resources from across the Institution, as reported during interviews.[30] The second looks at the descriptive systems and digital file storage systems in use by each collecting unit interviewed. We then discuss any gaps in preservation service amongst these combined systems.

[30] SIA Digital Archives, which is managed by SIA, is the one exception. Because SIA acquires content from across the Institution for its collections, as per its role as the archive of the organization, and because it is fairly significant in the volume and diversity of content it stores, its internal repository is included in this analysis.

2.2.1 Repositories

Interviews were conducted with stakeholders representing three cross-Institutional repository services as well as the SIA Digital Archives[31]:

Repository	Managed By	Serving
SI DAMS	OCIO, DAMS Branch	All Smithsonian museums, research centers, and offices that create digital image, audio, video, or time-based media art assets.
DSpace Digital Repository	SI Libraries	"All Smithsonian museums, research centers, and offices whose staff produce research publications are eligible to participate. The service will include data on publications authored by Smithsonian staff (federal and trust), staff from other agencies housed in the Smithsonian and working on Smithsonian collections, and affiliates, including research associates, graduate and post-doc students, and visiting scholars among others."[32]
SIdora	OCIO, Office of Research Information Services	Anyone engaged "in any research activity across the Institution, from Art History through Zoology"[33]
SIA Digital Archives	Smithsonian Institution Archives	SIA archivists

Table 4. Cross-Institutional repositories and stakeholders.

While none of these explicitly and formally function as preservation repositories, they often serve this role on behalf of users, in lieu of alternative solutions. As such, with increase in usage and volume comes the need for these repositories to provide more preservation services. While stakeholders are understandably confused by each repository's role with regard to preservation, submitting data to these repositories is better than leaving digital content on unmanaged storage devices and systems.

Given that these repositories currently are the only ones identified that attempt to answer preservation needs, it is useful to compare them in order to identify current and future gaps in service. Here we look at two parameters:

1. **Content and data type scope**. "Content" refers to the type of intellectual creation; "data type" refers to either the specific digital format or general digital media that the content is captured in.

[31] Note that this analysis does not include local business storage environments such as network shares within units.
[32] Smithsonian Research Bibliography Frequently Asked Questions. Smithsonian Institution. http://research.si.edu/srb_faq.cfm. Accessed September 28, 2016.
[33] SIdora Functional Overview. Smithsonian Institution. https://oris.si.edu/sidora-functional-overview. Accessed September 28, 2016.

2. **Storage volume** used, and total **storage capacity,** both as of June 2016.

All figures in Table 5 are as reported by interviewees.

Repository	Content Scope	Data Type Scope
SI DAMS	Archival content, artworks, collection items, communications materials, documentation of collections, educational production, event documentation, exhibition components, interviews, oral histories, research or study collections, time based media art (TBMA)	Video, audio, images, data associated with time based media art
DSpace Digital Repository	Publications and associated research data	No restrictions on file type, current content includes PDF, .epub, .mobi, .txt, .wav, .mp4, and tabular data
SIdora	Research data in an active state of creation or use	"almost anything" — 3D, GIS, text, spreadsheets, .pdf, audio, video, images
SIA Digital Archives	Born-digital archival holdings	A wide variety of formats, which may include: .pdf, spreadsheets, email, web (WARC), social media, databases, CAD, custom built software, blueprints, audio, video, images, raw data that scientists have collected, etc.

Table 5. Repository content and data type scope

Repository	# of files	Volume (June 2016)	Remaining Capacity (June 2016)
SI DAMS	8,821,627 files	1400 TB (x 2 copies)	600 TB (x 2 copies)
DSpace Digital Repository	20,000 files	0.39 TB	0.2 TB
SIdora	22,300,000 files	19 TB	11 TB
SIA Digital Archives	>750,000 files	16 TB	10 TB
TOTAL	31,891,627 files	1435 TB	621 TB

Table 6. Repository volume and capacity

SI DAMS is by far the largest and most mature repository of those examined. It stores and disseminates a variety of content, including collection items and their associated information

(metadata and other documentation), as well as Institutional output, such as educational and marketing materials. It does not address research data storage needs at this time.

Though it was developed initially for access purposes, SI DAMS has become the *de facto* storage environment for digital collection items and a large volume of Institutional output, and as a result, the need for it to operate as a preservation system has emerged. SI DAMS staff have taken this role seriously, and have added preservation functionality to the repository over time. As previously noted, AVPreserve's 2015 assessment of the DAMS found it largely compliant with the technical and procedural criteria of ISO 16363*,* with the exception of many required policies, which would need to be determined by Smithsonian senior management (see Section 1.1 Background and Previous Work).

It appears that parameters for defining both resource and functional scope of the various repositories varies. SI DAMS, for instance, has based its ingest requirements on data type (audio, video, images), although time-based media art is also in scope. Digital files are only stored in SIA Digital Archives if they are born-digital accessions (digitized collections held by SIA are stored in DAMS). SIL's DSpace repository is specific to publications and associated datasets, and SIdora's focus is on active research. From a functional perspective, SIdora does not intend to provide preservation services, but DAMS does. However, DAMS cannot necessarily accept data from SIdora or SIA Digital Archives for long-term preservation because it does not support those data types.

Submission policies and scope also appear to remain at the discretion of the repository management. Decision makers take responsibility for content within their scope, and feel the rest should be managed in a more appropriate environment for the content and/or usage. While in some ways this is a responsible sentiment (no repository should attempt to manage everything), there is no cross-repository oversight, coordination, or governance mechanism to ensure that all data that requires long-term stewardship has a long-term home. In particular, comparison of the systems reveals that there is very limited support today for research data between these repositories, both in storage capacity, and in research data types accepted. As is demonstrated in Section 3, the scale of research data may easily dwarf the storage capacity of all four repositories combined.

2.2.2 Systems in Use

Interviewees were asked to describe the various systems they use for management of data, including storage, description, access, and preservation environments. As the preservation of digital resources requires both digital files and the metadata and other associated information that surrounds them be part of the archival information package (AIP), any system that manages any part of this package was explored with the interviewee. This includes collection information systems (CIS), databases, archival management systems, digital asset management systems, storage environments, and other repositories.

Collections

The table (7) below summarizes only the CISs and digital asset storage and access systems used by the collecting units interviewed.

Collecting Unit	Collection Information System	Digital asset storage system
F|S	TMS, Archivists Toolkit	DAMS
NMAH	XG, Archivists Toolkit	DAMS
NMAI	EMu, Archivists Toolkit	DAMS
NMNH	EMu	DAMS
HMSG	TMS	DAMS, T: Drive (local)
CFCH	?	DAMS
SIL	SIRSI/Dynix Horizon	DAMS, Internet Archive, Isilon
SIA	CMS (local)	DAMS, SIA Digital Archives (local)

Table 7. Collection Information Systems (CIS) and Digital Asset Management Systems (DAMS) used by collecting units.

As is demonstrated here, all collecting units interviewed are using the DAMS for storage of and access to digital assets. Most, but not all units interviewed see the DAMS as a preservation system as well, and trust it to perform this function. A few interviewees feel that DAMS is a good service, but problematic because its scope and policies are determined by the DAMS Branch, a part of OCIO, which they feel should not be responsible for setting preservation policy, as they are a technical group, not a curatorial one.

Some collecting units are using other systems for asset management in addition to the DAMS. These are used in the case that the DAMS does not support a specific media type or data model (e.g., compound objects created from digitized texts held by SIL, or email archives held by SIA) or the unit feels that it is their responsibility to manage a local copy of the assets in addition to the DAMS copy (HMSG). Units also reported using external hard drives (HDDs) for management of some digital resources. While in many cases use of HDDs is only intended to be short term, it is not uncommon for these devices to become longer-term homes for digital resources. The files stored on these are at risk both due to the vulnerability of the storage device, and because they are less likely to be backed up and monitored.

The full information package is spread across multiple environments, including the CISs and digital asset environments. In many cases, units does not necessarily see the CIS data as a target of preservation. In other cases, the unit feels that the CIS data are critical assets in and of themselves, and require equal stewardship to the file-based assets.

Content Creators

There was less of a pattern in systems usage between content creators interviewed, given the variety of professions represented:

- Researchers are using a variety of storage environments, including: local storage such as hard disk drives, USB drives, and optical media; networked storage provided by SI; commercial cloud services such as Dropbox; storage offered by collaborating Institutions (e.g., universities); and third-party data repositories.

- Some research professions track and create metadata about their research output, but practice varies greatly. Spreadsheets are the most common way of tracking, although some specialties use specific descriptive systems (e.g., for frozen biological specimens).

- Facilities personnel use a variety of methods for managing Institutional output; shared file systems with consistent folder structures are the most common.

For more information on the technical environments used by researcher stakeholders, please see the results of the research data survey in Section 3.

2.3 Current Obstacles or Issues

Key question: What are the biggest challenges today that inhibit comprehensive digital preservation?

The 16 interview groups discussed a wide variety of issues they feel are an impediment to digital preservation at the Smithsonian. The specific examples raised were analyzed and distilled into 15 root or underlying causes. Table 8 lists these underlying challenges, with a total of how many groups raised the topic (out of 16), and the total number of distinct points made about this issue (i.e. one interview group may have raised three discrete issues that can be tied to the same root challenge).

ID	Root/Underlying Challenge	# groups that raised topic	# distinct points made
1	Expectations and responsibilities are not matched by staffing resources	14	26
2	Some important data types are not being systematically stewarded/collected/created	12	12
3	The central roles necessary for digital preservation to be enacted at an Institutional level are missing	9	10
4	Existing policies are insufficient or lacking in clarity to accommodate the digital landscape today	8	9
5	There are limitations in storage planning and capacity for digital assets	7	9
6	Researchers' incentives for data management are not aligned with the Institution's	6	10
7	The siloed nature of SI's units, policies, and systems is problematic	6	7
8	Roles and responsibilities with regard to digital preservation are often unclear and/or inappropriate	5	6
9	Several existing SI technologies are felt to be difficult to use and/or don't fulfill their function adequately	5	9
10	Proprietary and complex file formats in use today create unknowns about future renderability	4	4
11	There is no policy for research data management / preservation	3	3
12	There is a lack of clarity around terminology	3	5
13	Sometimes technology leads, rather than the users' needs	2	2
14	There is resistance to change to new digital norms	2	2
15	Documentation and communication is not always clear	1	1

Table 8. Underlying challenges to digital preservation by specific topic, organized by numbers of times topic was mentioned.

As is made clear by Table 8, the number one issue facing nearly all interviewees is a lack of resources with respect to existing Institutional expectations to meet digitization targets and resulting responsibilities, including digital preservation. Interviewees feel they are stretched in terms of time and staffing to meet existing mandates, and cannot confidently take on digital preservation as well. As will be discussed in Section 4 Findings: Document Review, current policies that relate to long-term management of digital resources do in fact put this burden on content creators (e.g., curators or collection managers who oversee digital preservation projects, or researchers).

Below, the top issues are further broken down by interviewee type — collecting unit, repository, content creator — in order to understand the different perspectives on these topics.

2.3.1 Collecting Units

The 8 collecting units interviewed plus the National Collections Program, as a representative of collecting units, are included in this analysis.

ID	Root/Underlying Challenge	# groups	# points
1	Expectations and responsibilities are not matched by staffing resources	8	17
2	Some important data types are not being systematically stewarded/collected/created	8	8
4	Existing policies are insufficient or lacking in clarity to accommodate the digital landscape today	6	7

Table 9. Underlying challenges to digital preservation by specific topic, organized by numbers of times topic was mentioned by Collecting Units.

1 — Expectations and responsibilities are not matched by staffing resources

Mentioned by 8/9 interview groups, with a total of 17 distinct points raised

Interviewees note that the primary factor contributing to the rapid increase in digital resources within collecting units is the mandate to digitize collections. However, they almost unanimously feel that there are not adequate staff at the unit level to meet this charge. A particular concern expressed is that units cannot adequately fulfill the metadata requirements that go along with digitization, with one group noting that, “metadata is the biggest problem by far." Their concerns are primarily around descriptive and rights metadata creation to support access — this doesn’t even account for preservation. They feel that adding digitization, not to mention digital preservation, to existing responsibilities is too much for existing staff to take on.

Interviewees also talked about the challenge of not having a dedicated point person within the unit to set internal procedures for digital resource management. Those units with this staffing role have relatively efficient and clear workflows. Those without it find that digital resources may be neglected, or they have unnecessary duplication of effort and inconsistent practices. Interviewees also expressed frustration that there are not enough staff in the DAMS Branch to provide ongoing support to all units.

2 — Some important data types are not being systematically stewarded/collected/created

Mentioned by 8/9 interview groups, with a total of 8 distinct points raised

While collecting units are concerned about the preservation of digital collection items in general, they feel that there are particular types of content that are not being stewarded at all. Research data (e.g., field notes) and institutional output (including podcasts/webcasts, exhibition

materials, email, and business records not collected by SIA) were mentioned as examples. Interviewees note that there is a significant amount of overlap and gray area between the collections, research data, and institutional output, and a frustrating degree of policy ambiguity, which further exacerbates the problematic stewardship of the various data types.

Some interviewees also remarked that certain types of metadata, particularly preservation metadata, was not being systematically collected or generated so that it can be used for purposes such as obsolescence monitoring.

4 — Existing policies are insufficient or lacking in clarity to accommodate the digital landscape today

Mentioned by 6/9 interview groups, with a total of 7 distinct points raised

In general, interviewees feel that existing policies do not adequately address digital preservation needs. SD 600 is felt to be broad enough to apply to digital collection items and associated information, and yet interviewees seemed unclear how to implement this directive for their unit's digital collections. For some, this was because the terminology is felt to be outdated with regard to digital information. Furthermore, units felt that SI's Unit Digitization Plans and Digital Asset Management Plans, mandated by SD 610, do not go far enough to truly ensure that the resulting digitized data is properly described, documented, or cared for over the long term: they leave decisions to project managers, there is little accountability, and the follow through on these plans is not always performed.

2.3.2 Repositories

This analysis is based on input of the 4 repository services interviewed.

ID	Root/Underlying Challenge	# groups	# points
1	Expectations and responsibilities are not matched by staffing resources	3	5
3	There are missing central roles necessary for digital preservation to be enacted at an Institutional level	3	3
5	There are limitations in storage planning and capacity for digital assets	3	4
6	Researchers' incentives for data management are not aligned with the Institution's	3	4

Table 10. Underlying challenges to digital preservation by specific topic, organized by numbers of times topic was mentioned, by Repository Service.

1 — Expectations and responsibilities are not matched by staffing resources

Mentioned by 3/4 interview groups, with a total of 5 distinct points raised

Staffing shortages limit how much outreach and support repositories can provide to units and researchers. Interviewees have found that education, training, and general communication go a long way toward raising awareness of repository services and helping people use those services effectively, though they are often unable to perform these functions on top of pressing, day-to-day management tasks.

Repositories have also found that units and researchers will often keep digital data on local storage (e.g. external hard drives) for extended periods of time. When they do finally hand it over, the work it takes to sort through all the legacy material, target original files, track down metadata, and archive the dataset is monumental. Multiple groups noted that the staff they have is inadequate compared to the volume of data they receive and the unique needs of the different datasets.

An interesting point was raised by the DAMS team that when collecting units take responsibility for their own archived content, which they often store offline on data tape, the tapes are not checked regularly due to limited people, time, and tools required to do so. This points to the need for increased adoption of central repository services, which can take on the responsibility of ensuring that data integrity is managed on behalf of stakeholders and are more likely to have the resources to perform the underlying risk management functions of digital preservation.

Increasing staff to provide improved repository services is reported to be problematic. There was a recognition that staffing is the most expensive part of the problem and also the most difficult to get. Hiring contractors is seen to be easier but more expensive.

3 — There are missing central roles necessary for digital preservation to be enacted at an Institutional level

Mentioned by 3/4 interview groups, with a total of 3 distinct points raised

Three of the four repositories stated unequivocally: there is no oversight for digital preservation at the Smithsonian. This may seem more obvious to the repositories than the collecting units, as these groups have greater visibility across the Institution than the individual units do, and/or because they themselves have nowhere to turn for guidance on preserving the digital assets in their care.

5 — There are limitations in storage planning and capacity for digital assets

Mentioned by 3/4 interview groups, with a total of 4 distinct points raised

Repositories find it challenging to scale up storage capacity to support a growing volume of submitted data because storage planning tends to be reactive rather than proactive. In at least one case, “the budget for storage is flat and does not accommodate the increase in born-digital accessions coming through the door.”

A separate but related point was raised by one repository that has observed loss of data at the unit level due to retention schedules for network drives that only go back one year. This points

to the need to avoid using local network shares for long-term storage, and increasing the storage capacity of repositories.

6 — Researchers' incentives for data management are not aligned with the Institution's

Mentioned by 3/4 interview groups, with a total of 4 distinct points raised

It was noted by several interviewees who frequently work with researchers that their incentives for preservation are much different than those of the Institution. First, it was noted that they are often more loyal to their field of study than they are to the Institution, and therefore will sometimes submit data (primarily for publication but with the assumption that long-term management will be a part of the service) to external services, which may be either a domain-specific repository, or a partner university. Doing so has the potential to help them go to where they might receive the most recognition and citations, as well as helps them contribute to further the knowledge of their field. Decisions have not yet been made at an Institutional level as to what the implications are for research data being managed externally, and what SI's role with regard to the data is once it leaves.

Interviewees remarked that scientists are interested in ensuring that their data and their legacy is preserved, and will do whatever they can to get it managed, particularly as they approach retirement. Therefore, if a university's offer is more attractive than what the Smithsonian can offer, they will deposit their data with that organization.

2.3.3 Content Creators

This analysis includes a total of four groups or individuals whose digital preservation responsibilities are largely focused on content creation — 2 researchers, the 3D imaging team, and the facilities Technology Advisory Council — as well as the perspective of the Pan-Smithsonian Cryo Initiative, which works extensively with researchers across the Institution, for a total of 5 interview groups.

ID	Root/Underlying Challenge	# groups	# points
1	Expectations and responsibilities are not matched by staffing resources	4	5
3	There are missing central roles necessary for digital preservation to be enacted at an Institutional level	4	4
9	Several existing SI technologies are felt to be difficult to use and/or don't fulfill their function adequately	3	7

Table 11. Underlying challenges to digital preservation by specific topic, organized by numbers of times topic was mentioned, by Content Creators.

1 — Expectations and responsibilities are not matched by staffing resources

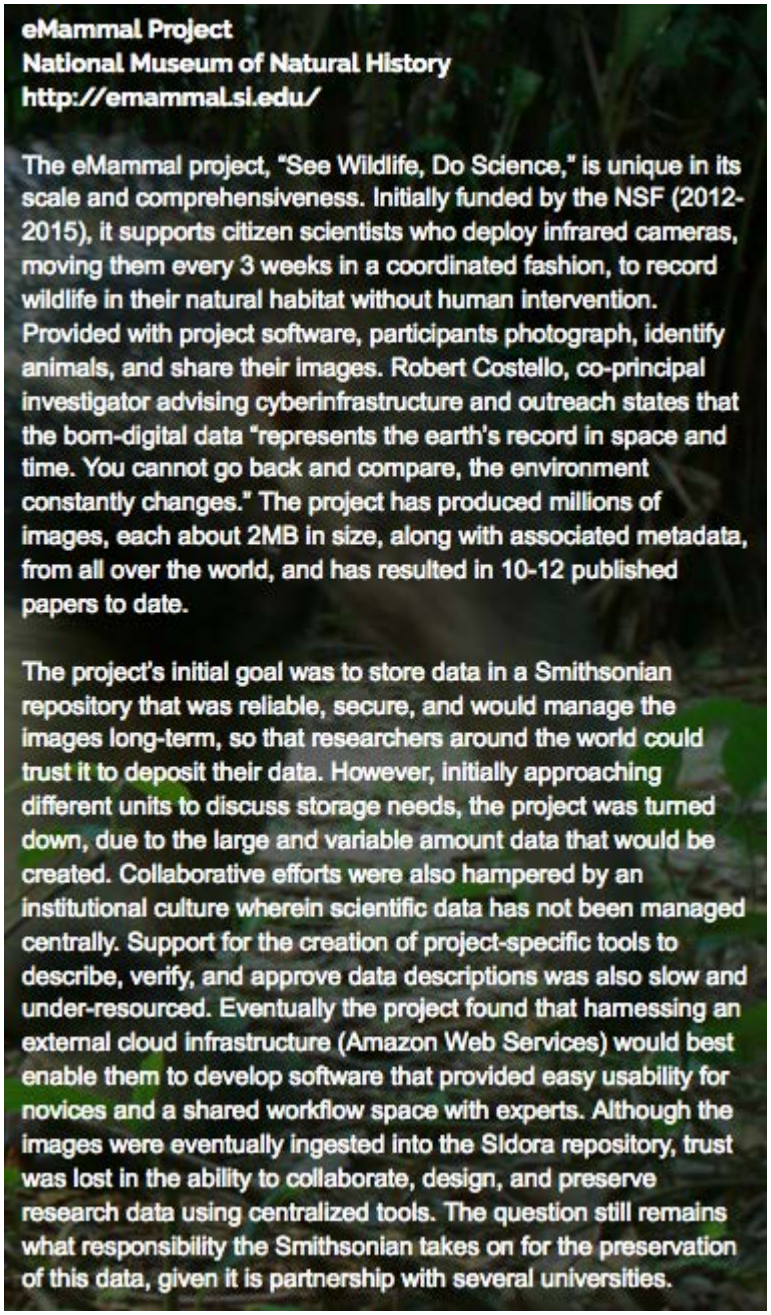
Mentioned by 4/5 interview groups, with a total of 5 distinct points raised

From the perspective of content creators, this challenge relates to three issues:

1. Researchers are busy, and don't want to be asked to do more than they already do in order to use a repository to manage their data
2. Smithsonian repositories don't have the staff resources to support researcher's needs
3. When research departments receive funding for data management, they don't always have the staff or know how to use it effectively.

3 — There are missing central roles necessary for digital preservation to be enacted at an Institutional level

Mentioned by 4/5 interview groups, with a total of 4 distinct points raised



Content creators, particularly researchers, feel that there is no central support service from the Smithsonian to help with ongoing management, preservation, and accessibility of research data. They note, however, that this isn't surprising given that there is no tradition of centrally managing research data at the Smithsonian. Interviewees see elevating the status and need for research data management is a chicken and egg problem: there is no central voice to advocate to senior management on behalf of researchers needs, and yet the needs must be made more visible in order for a central voice to emerge.

Beyond researchers, a related issue was voiced by the 3D team, specifically that content creation is being performed by both central offices as well as units, yet there is no guidance for these disparate groups to ensure that they create high-quality, archive-ready scans.

9 — Several existing SI technologies are felt to be difficult to use and/or do not fulfill their function adequately

Mentioned by 3/5 interview groups, with a total of 7 distinct points raised

This was a very important topic for several groups interviewed, particularly researchers who have been working with Smithsonian services on various relevant aspects of data creation and management. They voiced several frustrations regarding technologies available at SI, citing issues such as:

- “The DAMS is like a deep freeze.” For active research, Dropbox and Google Drive work better, as they are easy to use and globally accessible.
- SIdora is one offering that SI makes for active research data management. However, users complain that SIdora requires that they, “learn a whole new way of organizing data.” They prefer simple, familiar tools. In fact, SIdora did not meet the needs of one large project, so they ended up building their own custom solution. One SIdora user said he, “would not use SIdora again.”
- There is no repository to provide full functional support for specific data types, like 3D or gene sequences. Either the formats are not supported, or the requisite computing resources are not available.

See Section 3 Findings: Researcher Survey for more detail on technical environments used by SI researchers from various disciplines.

In conclusion, top concerns were generally shared by all three types of stakeholder groups interviewed: digital preservation is more than most can take on given their existing responsibilities and mandates, yet they feel strongly that the digital resources in their care require a great deal more stewardship than they are currently afforded. They point to the lack of central roles as a gap at the Institution-level that challenges the current situation.

Researchers clearly feel challenged by the concept of data management. While they are federally mandated to ensure the long-term availability of their data, they are not always finding the guidance and infrastructure support they need within the Smithsonian to fulfill these requirements. Because grant guidelines leave decision making up to the researcher, it would be up to the Smithsonian to mandate that the data that is generated within the Institution be managed in a specific way. Without any formalized position on this, the Smithsonian doesn’t have much say over what happens to the research data produced within the Institution, and there are liabilities if the data is not managed for the duration of the grant period and beyond.

2.4 Ideas and Opportunities for Improvement

Key question: What changes could be made to improve preservation of digital resources at the Smithsonian?

It was clear that interviewees care deeply about the digital resources they are responsible for creating and or managing, and ensuring the longevity of content is a top priority for each stakeholder. Interviewees expressed wide range of ideas for improvements that can be made to

better steward these resources over time that come from their experience and knowledge of the keys to a successful program at the Smithsonian.

ID	Idea for Improvement	# groups that raised topic	# distinct points made
1	Establish centralized digital preservation services & infrastructure	8	10
2	Establish shared responsibilities between content creators, units, and central services	8	12
3	Improve/enforce policies to better support digital preservation	6	6
4	Increase staffing at both the central and unit level to support preservation	5	6
5	Provide guidelines and training to creators and collecting units	4	5
6	Demonstrate success in order to make a solution attractive	3	4
7	Look at funding models to support digital preservation	3	3
8	Look to other industries which may have made more progress on the digital preservation issue	3	3
9	Establish committees to advise and guide a digital preservation program	3	3
10	Prioritize content / stop the bleeding	3	3
11	Establish and enforce a digital preservation mandate	3	3
12	Create an easy-to-use preservation infrastructure	3	3
13	Standardize terminology to establish common reference points for digital preservation and support implementation	3	6
14	Offer incentives for preservation	3	3
15	Capture data at point of creation	2	2
16	Create stewardship policies for research output	2	2
17	Define the needs and requirements, and create a plan of action	1	2
18	Establish classes of content for different levels of care	1	1

Table 12. Ideas for improvements that can be made to better steward digital resources over time by specific topic, organized by numbers of times topic was mentioned.

Several of the top suggestions are further explored below. These and others will be revisited in Section 6 Recommendations, later in this report.

1 — Establish centralized digital preservation services & infrastructure

Mentioned by 8/16 interview groups, with a total of 10 distinct points raised

This suggestion was made by interviewees in each category. Taken together, their suggestion can be summarized as follows:

- There should be a digital preservation directorate that can set policy, provide guidance, and advocate to Senior Leadership. The directorate should determine roles of SIA, SIL, OCIO, units, and content creators. The National Collections Program was mentioned by multiple interviewees as a good model for this.

- There should be an easy-to-use, common preservation infrastructure to support a digital preservation mandate. This would create efficiencies, reducing the need for everyone to maintain their own environments. CollectionsSearch and DAMS were mentioned as good models of existing central technologies.

2 — Establish shared responsibilities between content creators, units, and central services

Mentioned by 8/16 interview groups, with a total of 12 distinct points raised

While it was argued that, “people understand the value of centrality,” interviewees also underscored the fact that preservation must be a shared responsibility between central services, and the units and content creators themselves. One group characterized the role of this relationship as, “Trust but verify," specifying that the unit should ultimately be responsible for preservation, but work with a central service to achieve it.

It was also suggested that digital preservation should be supported by an ecosystem of technologies, not one system. DAMS integration with the CISs was offered as a model for repository integration. However it was emphasized that there must be an identifiable repository for all resources of value, which provides functionality that meet stakeholders’ requirements.

3 — Improve/enforce policies to better support digital preservation

Mentioned by 6/16 interview groups, with a total of 6 distinct points raised

This overarching recommendation encompasses several specifics, which include:

- Create new policies that would support a digital preservation mandate, including policies for research data
- Update SD 600 to make it more relevant to digital resources
- Revisit SD 609 and SD 610 and emphasize the management and digital preservation components or create a new directive for this
- Enforce existing policies: make units and content creators accountable

4 — Increase staffing at both the central and unit level to support preservation

Mentioned by 5/16 interview groups, with a total of 6 distinct points raised

Interviewees emphasized that in order for a shared/central model to be successful, more staffing resources would be needed across the board. At the unit level, especially for the larger units, an in-house digital asset manager could set internal procedures and act as a liaison to central services. Within central services, such as DAMS, additional staff are needed to provide dedicated support to units.

Interviewees also remarked that it will be critical for there to be people who can provide help to researchers, or they won't know what to do, and follow through on any new Institutionally mandated data management requirements would not be likely.

5 — Provide guidelines and training to creators and collecting units

Mentioned by 4/16 interview groups, with a total of 5 distinct points raised

This suggestion was raised in the context of both collections and research data. For collections, the emphasis was on providing guidance that units could use for implementation of policies with regard to digital assets and associated information. It was also suggested that curators should have training so that they will have a better understanding of the factors that go into long-term access of digital information and what the risks are.

For researchers, it was stressed that clear guidance is needed from the point that they are hired or that their fellowships begin: what the Institution expects them to capture, and where, when, and how they should deposit it. The suggestion was also made that templates for data management plans be provided, and support throughout the researcher's career be available, to help them best prepare for retirement.

13 — Standardize terminology to establish common reference points for digital preservation and support implementation

Mentioned by 3/16 interview groups, with a total of 6 distinct points raised

Although this suggestion was only made by a handful of interview groups, those who did bring it up felt very strongly about the need for clarity around digital preservation-related terminology, as they feel this is a critical underlying cause of confusion and complacency when it comes to coordinating efforts today. Specifics included:

- Update SD 600 to eliminate outdated terminology and add contemporary language when referencing digital resources
  Strengthen concept of "associated information" in SD 600 for a digital world
- Formalize and communicate all relevant terminology
- Establish a definition of "digital preservation" for the Smithsonian Institution

Many of the suggestions provided by interviewees were factored into the recommendations provided in this report, which can be found in Section 6.

3 FINDINGS: RESEARCHER SURVEY

One of the original goals of this project was to produce quantitative information regarding the scale and types of digital research data that is produced in and held by the Smithsonian currently, and to provide projections for the near future.

In speaking with the Digital Preservation Working Group and stakeholders who participated this study, and in reading past the reports of past studies, it is clear that for some time there have been questions raised and discussions held addressing the extent to which, and how the Smithsonian might, address digital preservation of research data. These questions are distinct from and often challenged by unanswered questions around how much and what kinds of data there are. The overwhelming nature of the unknowns have made any theoretical or pragmatic progress difficult, resulting in inaction. This report is a major leap forward in beginning to address these issues.

The interviews conducted in spring and summer 2016 for this study yielded valuable insights, but were limited in how much they could quantify the landscape. It was determined that a survey aimed at active Smithsonian researchers would be a useful instrument to gather more in-depth information about how, how much, and types of data are being produced at the Institution. The survey was open from August 4th through the 22nd.

Survey Questions

- What is your field of study? Do you work for a specific SI Museum unit/department and/or research center/program?
- Where do you store research data?
- How much total digital research data have you created and store on the systems described above? What percentage of this total is primary / raw, analyzed / derived, and publication data?
- What do you estimate your data holdings will be in 5 and 10 years?
- Which types of research data should be retained for future re-use and/or reproducibility (primary / raw, analyzed / derived, and publication)? For what period of time?
- If you have shared research data, how have you done so?
- Do you use any specialty data formats that are common to your field for data sharing?

The survey was designed to be concise and focus on gaining insights into:

- Total volume of data holdings today, and how researchers anticipate their data storage needs to grow in the coming years
- Volume and retention of data by type
- How researchers are storing their data
- If and how researchers are sharing their data

This information makes the implications for and risks to data more clear throughout SI. Through this survey, challenges of digital preservation are made real — the resources required become clearer and the volume and types of data at risk of loss are quantified. Even though the responses are sometimes overwhelming, having actual numbers to work with can only help inform decision making and spur progress to address real needs.

The survey captured 100 complete responses from researchers across multiple domains, which were grouped together according to the following:

Group	Abbreviation	# Respondents
Museum Conservation Institute	MCI	2
Nation Air and Space Museum	NASM	6
National Museum of Natural History	NMNH	35
Genomics	Genomics	3
Smithsonian Astrophysical Observatory	SAO	34
Smithsonian Conservation Biology Institute, Smithsonian Environmental Research Center, Smithsonian Tropical Research Institute	SCBI-SERC-STRI	20

Table 13. Number of survey responses by group.

All of the raw survey data and a breakdown of each of the responses can be found in **Appendix E. Survey Data - Raw**. This section focuses on the results of the survey, the conclusions that can be made from it, and other takeaways that serve to improve the understanding of the current (and future) status of data at the Smithsonian. With a better understanding of researcher data, the Smithsonian can begin to consider how to actively manage data now and forecast growth for expanding services to these stakeholders in the future.

3.1 Data Amounts

3.1.1 Data Total Calculation

The total data reported in this survey across all data types and all storage locations was 623 TB. This averages out to 6.23 TB per respondent for 100 respondents. During the interviews conducted for this study, we received estimations that there are somewhere between 850 and 1,250 researchers within the Institution, with the caveats that there are researchers of different types (staff scientists, curators, research fellows, interns, etc.), that play different roles, and varying length of service to the Institution. With these caveats in mind, it is reasonable to use 1,000 as the number of researchers within the Smithsonian with which to extrapolate from the survey data gathered. While it may seem obvious to multiply the average per researcher by 1,000 researchers we feel that it is in fact best to first average by person per domain before multiplying by 1,000. This is important because of the great and meaningful variances between domains, shown in the chart below.

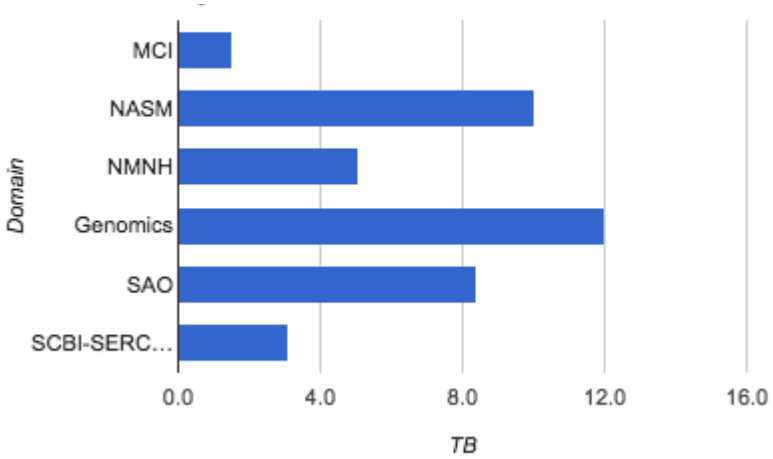


Figure 2. Data generation by domain, based on current average TB per researcher per domain

Calculating the average TB per person, per domain results in approximately **6.7 TB per researcher**. This brings the extrapolated total to 6,700 TB, or **6.7 PB.** However, another important consideration of the survey is that the sample size is seemingly small for a few of the domains. Notably, Genomics, which represents the the largest producer of data per response, only has a sample size of 3, and MCI, the smallest producer of data, only has a sample size of 2. One way to address this would be to identify the number of researchers per domain and then extrapolate based on the percentage represented by the sample size for each domain in order to come to a total. However, based on the analysis performed and in the interest of having a logical number to work with, we believe that averaging per person by domain is currently the most meaningful way to arrive at a data generation baseline number, making the point of reference used in this report 6.7 PB.

3.1.2 Future projections

To gain insights into the future data generation totals we asked researchers how much total data they anticipated having generated in total by the years 2021 and 2026. The survey results lead to projections that the amount of data will grow 4.7 times by 2021 and 41.8 times by 2026, reaching an estimated 280 PB across all researchers over the next 10 years.

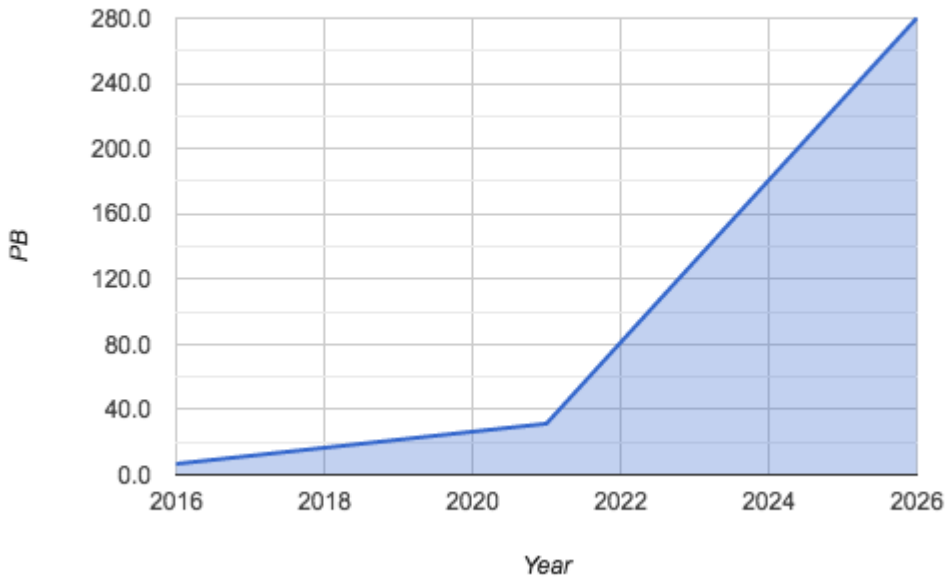


Figure 3. Research data growth through 2026.

3.2 Data Types

In the survey, researchers were asked about the types of data they produce and will need access to over time. Data types were analyzed by life cycle type, which are broken down according to the following definitions:

- **Raw / Primary data:** Defined for the purpose of this survey as data generated through observations, instruments, and/or experiments, either as originally collected or after being checked, calibrated, and/or organized (e.g., survey responses, sensor data, neurological images).

- **Analyzed / Derived data:** Defined for the purpose of this survey as refined data derived from Raw / Primary research data and interpreted by the researcher through some form of manipulation, transformation, or abstraction (e.g., statistical analysis or modeling).

- **Publication data:** Defined for the purpose of this survey as reference or canonical data that is a subset of Analyzed / Derived data and is likely peer reviewed, published, and/or curated. Publication data is potentially a synthesis of ideas and datasets from several researchers.

Breaking the data amounts down by life cycle offers a meaningful way to split up the data for thinking about varying retention periods, requirements, use cases, tiers of storage and more. The chart below shows the amount of data by life-cycle stage.

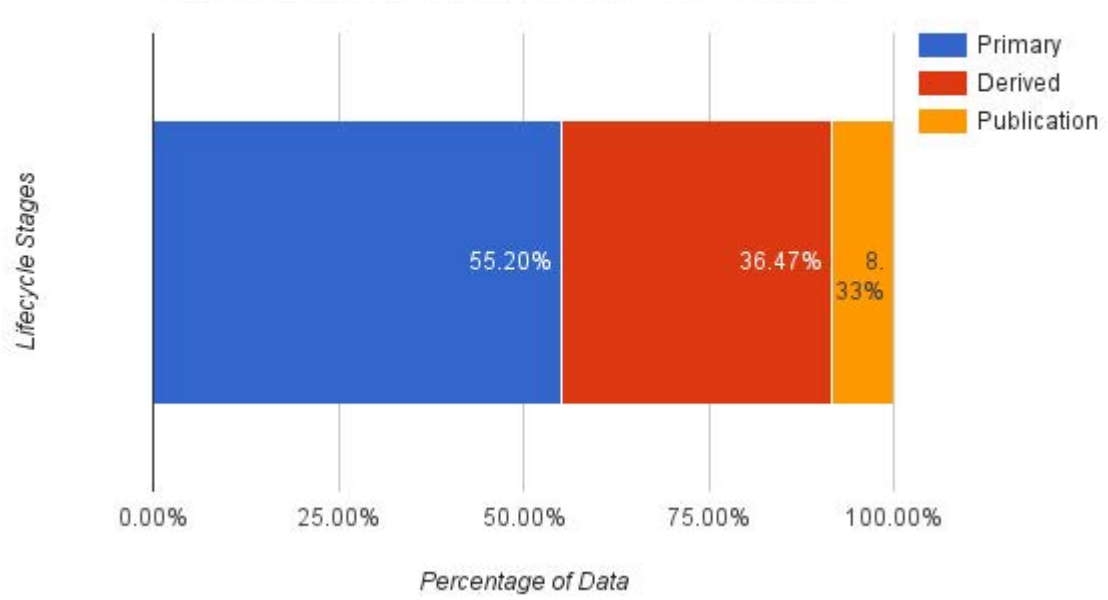


Figure 4. Data amounts by life cycle across all domains

Averaging by respondent within each domain, the findings vary slightly.

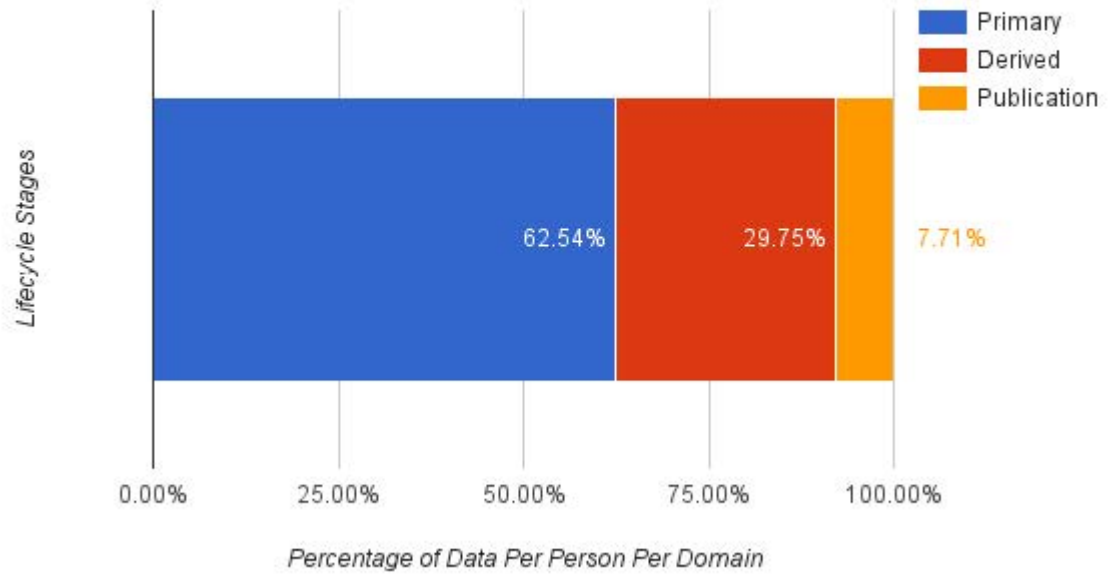


Figure 5. Data amounts by life cycle per person per domain

Using the latter percentages and calculating against the 2016 and 2026 estimated totals we see the following allocation across these data types:

2016
Raw / Primary data: 4.19 PB
Analyzed / Derived data: 1.99 PB
Publication data: .52 PB

2026
Raw / Primary data: 175.1 PB
Analyzed / Derived data: 83.3 PB
Publication data: 21.6 PB

3.3 Data Retention Periods

Retention periods are an important factor because they speak to how much of the total data needs to be kept for how long, and they inform the infrastructure and resources necessary over time. Respondents collectively identified the following average retention periods across all data types:

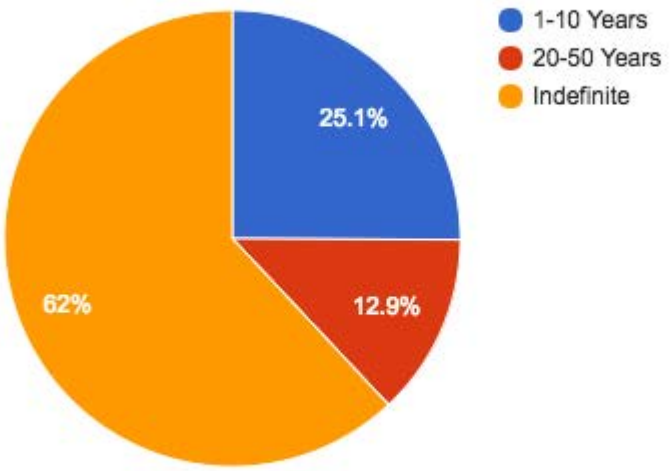


Figure 6. Retention periods for all data

Looking at the retention period across all data over the next 10 years based on these results provides the following outcomes:

Retention Period	2016 (PB)	2021 (PB)	2026 (PB)
Indefinite	4.10	19.16	171.19
20 - 50 years	0.85	3.99	35.62
10 - 20 years	1.66	7.75	69.27

Table 14. Data growth by retention period, 2016-2026

The survey also asked researchers to identify the most accurate retention period for each life cycle stage. A few either did not answer the question or stated that no retention period was necessary (for instance, for derived data, assumedly because it could be derived again from the primary data) so the total percentages do not add up to 100%.

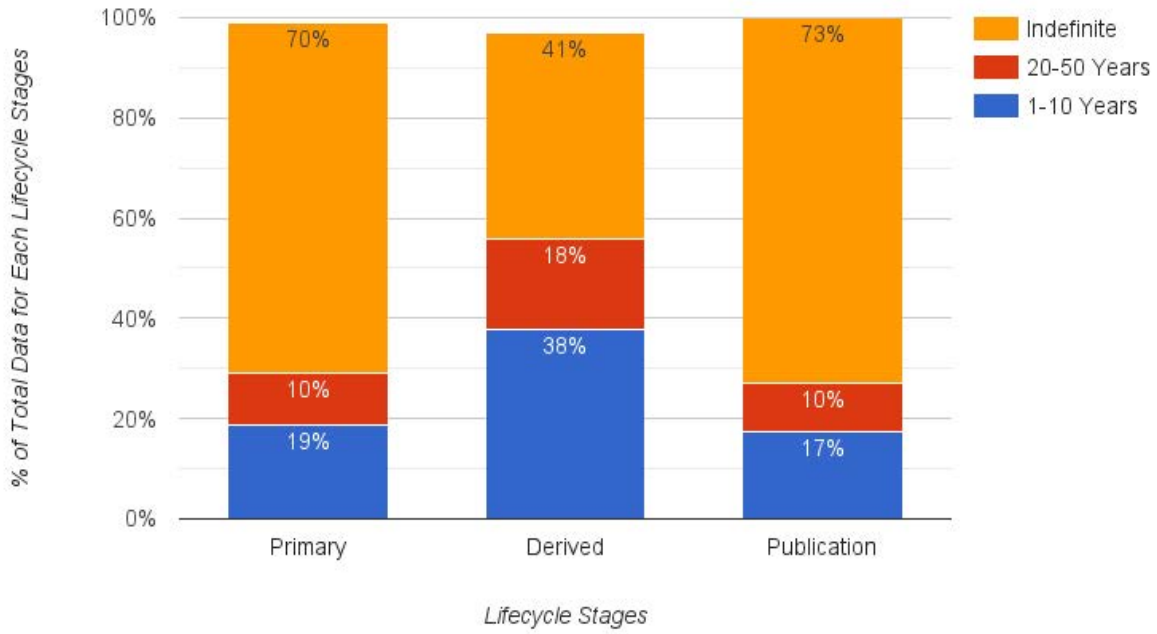


Figure 7. Retention period by data life-cycle stage

Based on 2016 and 2026 data totals this results in the following breakdown:

2016

Retention Period	Primary (PB)	Derived (PB)	Publication (PB)
Indefinite	2.58	1.00	0.41
20-50 Years	0.39	0.44	0.05
1-10 Years	0.69	0.93	0.10

2026

Retention Period	Primary (PB)	Derived (PB)	Publication (PB)
Indefinite	107.86	41.92	16.99
20-50 Years	16.10	18.27	2.28
1-10 Years	28.98	38.70	4.06

Tables 15 and 16. Data totals by retention periods per life-cycle stage, 2016 and 2026.

Reviewing the allocation of data by life-cycle stage and retention period begins to hint at the potential for tiered approaches and systems, addressing each of the data types and retention periods with its own set of policies and practices. While there are additional questions that would offer useful information, such as the frequency and type of access required of each data type, this provides a good start into thinking about how to divide, prioritize and create a phased/tiered approach.

3.4 Data Storage Locations

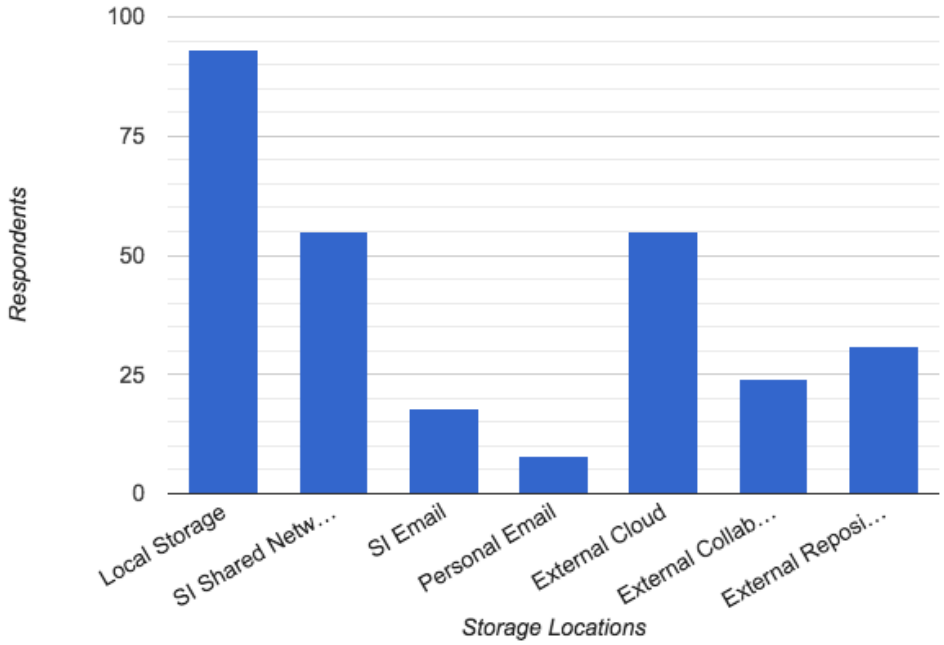


Figure 8. Data by storage location

The graph above shows how many respondents stated that they store data in a given storage location. Nearly all respondents reported that they store data from all life-cycle stages on Local Storage (e.g., removable hard drives, DVD, CD).

As seen in the graph below, SI Shared Network, External Cloud, and External Collaborator storage locations skewed toward storage of primary and derived data, while respondents heavily favored External Data Repositories for published data. Email Storage, not pictured below, is used fairly equally for all data types.

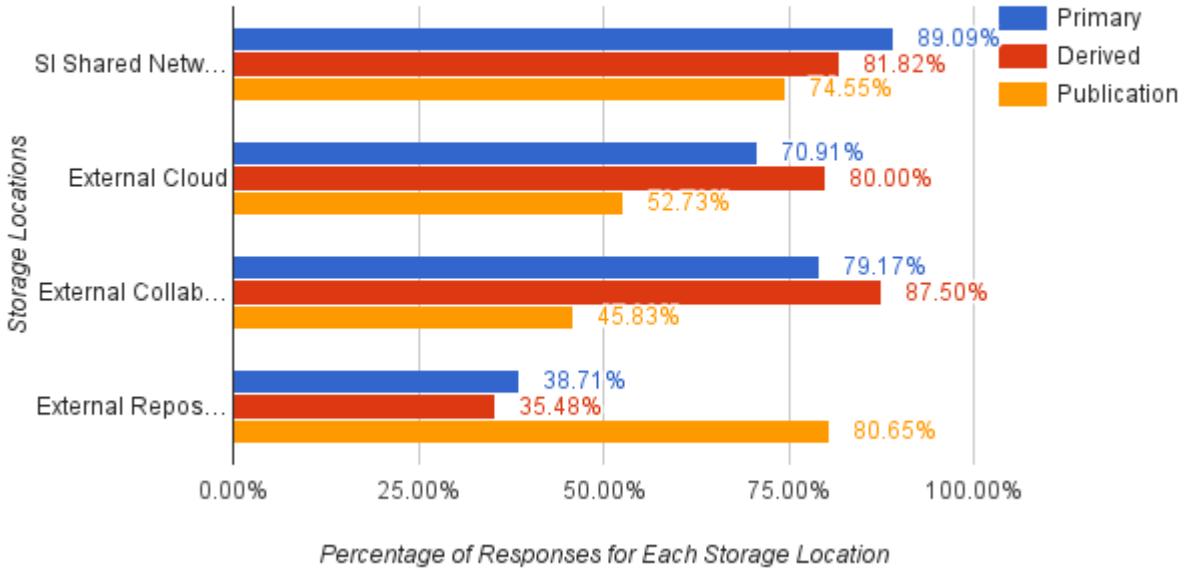


Figure 9. Storage locations by data life-cycle stage

One insight that this analysis provides is the amount of data that is in a managed environment compared to an unmanaged environment. For the purpose of this report, a managed environment is defined as an environment which is governed by preservation policies and practices. An unmanaged environment does not have preservation policies or practices and therefore the data stored within them is at increased risk of loss.

Breaking the storage locations into managed and unmanaged environments is the first step in performing this analysis. External Data Repositories fall within the managed category. Collaborative Partners may or may not be managed environments. For the purpose of this analysis we assume that 50% of these are managed. Within the SI Shared Network, the most widely-used managed environment for research data is SIdora, which to date has only supported a handful of projects. Given the small amount of data represented by this handful of projects and the relationship in size between SIdora and the total size of the SI Shared Network, for the purpose of this analysis we are considering the entirety of the SI Shared Network to be unmanaged. No other storage locations are considered managed environments from a preservation context. Note that the DAMS is not factored into the storage location analysis because it does not contain research data.[34]

We do not know the amount of data that is stored in each storage location. In the absence of explicit data, we can use as an allocation based on the relative number of responses for each storage location. Respondents were allowed to select all storage locations that they use for each life-cycle stage of data. Comparing the total number of responses for each storage location against the overall total number of responses provides us with the following distribution.

[34] Per input from DAMS team during interviews.

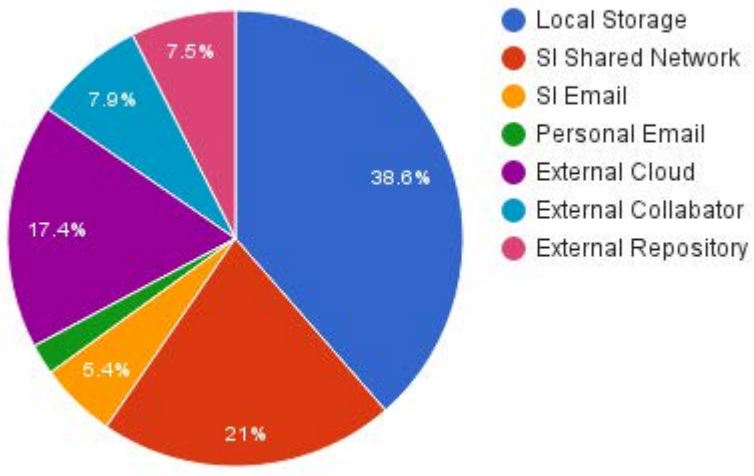


Figure 10. Storage locations based on percentage of responses

Using these numbers, we can extrapolate the 2016 and 2026 total data storage projections per location:

Year	Local Storage (PB)	SI Shared Network (PB)	SI Email (PB)	Personal Email (PB)	External Cloud (PB)	External Collaborator (PB)	External Repository (PB)
2016	2.58	1.41	0.36	0.15	1.17	0.53	0.50
2026	107.99	58.79	15.24	6.10	48.77	22.21	20.90

Table 17. Total data storage per location, 2016 and 2026

Using these numbers, and incorporating the stated logic regarding managed and unmanaged environments, we arrive at the following numbers for the current data totals.

	Managed	Unmanaged
PB	.77	5.93
% of data	11.5%	88.5%

Table 18. Current data totals, managed and unmanaged storage

Another way to look at this is to identify the respondents who store their data in managed environments and then apply the allocation seen in Figure 10. We found that 49 respondents store data in in managed environments. If each researcher represents 6.7 TB of data, and we use Figure 10 to assume that 7.5% of their data is in an External Repository and 19.3% is in a managed External Collaborator's environment (given 50% of External Collaborator storage is assumed to be managed), the total comes out to 1.8 TB per researcher, or 882 TB across the 49 identified researchers that store their data in managed environments. 882 TB represents approximately 13.2% of 6.7 PB, resulting in the following breakdown:

	Managed	Unmanaged
PB	.88	5.8
% of data	13.2%	86.8%

Table 19. Current data totals, managed and unmanaged storage (alternate view)

Both methods of analysis produce results within 2% of each other, ranging from 86.8% to 88.5% of data being unmanaged. The proximity of these two results, created through two different methodologies, provides some confidence in the absence of explicit information regarding the amount of data stored in each location. For the purpose of this report we will use the the smaller of the two numbers, and the resulting current unmanaged data total of 5.8 PB.

This analysis also reveals that a great deal of Smithsonian research data is not controlled by the Institution in any way because it is not stored on a Smithsonian centrally managed system. Using Figure 10 above, we see that only 65% of responses pointed to storing data internal to the Smithsonian (Local Storage, SI Shared Network, SI Email), and only 26.4% pointed to storing data on Smithsonian centrally managed systems (SI Shared Network, SI Email). The lack of control by the Smithsonian (and absence of preservation measures currently taking place) for the vast majority of data being generated by researchers is a significant potential risk for the loss of data. And, it means that the speed and difficulty by which the Smithsonian will be able to gain control over and governance of its research data will be tremendous, should it decide to do so.

4 FINDINGS: DOCUMENTATION REVIEW

An examination of current working strategic plans, policies, and previous analysis that document the Smithsonian's commitment toward preservation of digital resources helps situate the findings revealed by the interview and survey processes within the Institution's existing strategic and operational goals. This review focuses on a handful of key documents that have the greatest impact on the day-to-day management of digital resources:

- Smithsonian Institution Digitization Strategic Plan, Fiscal Years 2010-2015
- Smithsonian Directive 600 — Collection Management (2001)
- Smithsonian Directive 610 — Digitization and Digital Asset Management Policy (2011)
- Sample Digital Asset Management Plans (DAMPS) (2013)
- NSF guidelines for Data Management Plans (DMPs)
- Sharing Smithsonian Digital Scientific Research from Biology (2011)

A review of relevant documentation reveals that there is a broad recognition that digital preservation is an important and integral part of the Institution's mission and digital ambitions. It is further recognized that successfully ensuring that digital assets are made available today and remain available in the future requires pan-Institutional coordination as well as new program development. However, we also find several shortcomings in the existing documentation with regard to how digital preservation is executed, which we feel contributes to or exacerbates the challenges expressed by the interviewees, and the issues highlighted by the survey results.

4.1 Digitization Strategic Plan

The Smithsonian's commitment to digital preservation is identified in the first goal of the 2010-2015 Digitization Strategic Plan: "Goal 1: Digital Assets — Provide unparalleled access to Smithsonian collections, research, and programs by creating, managing, and promoting the Institution's digital assets," specifying that, "We must establish trusted digital repositories to preserve the assets once digitized, and then ensure that we can integrate them across the Smithsonian and into the broader online arena."[35]

While the need for digital preservation is clearly recognized in this foundational document, there is a lack of specificity on what actionable steps should be taken toward the implementation of Institution-wide digital preservation, especially when compared to the detail to which digitization tasks are outlined. In fact, of the three goals laid out in the plan, digital preservation is only addressed once, in Objective 1 of Goal 1, which directs staff to, "Protect and enhance the value of all Smithsonian digital assets through coordinated digital asset management," detailing that there be an effort to "develop requirements for life-cycle management of digital assets to ensure

[35] Ibid., 11.

immediate access and long-term preservation.[36] This is just one of 17 tasks across the three Objectives of Goal 1. Goals 2 ("Digitization Program — To pursue its mission in the 21st century, integrate digitization into the core functions of the Smithsonian.") and 3 ("Organizational Capacity — Through novel, innovative approaches, secure sufficient resources and build capacity to create and sustain a digital Smithsonian.") do not mention digital preservation at all, and are exclusively focused on the digitization and access efforts.

The preservation-related task prescribed by this document is an important one; developing requirements for life-cycle management is a critical step toward effective stewardship of the Institution's resources. Such an activity could look broadly at research data, digital collections, and more, identifying unique requirements of stakeholders in different arenas. But it is also just that, a step. The plan does not identify what further steps should be taken beyond this one, what kinds of programmatic developments should be made, or human, organizational, or technical resources should be devoted to this objective. Reading it now several years later, it feels like a vague commitment, and begs the question, can it be confidently stated that this step was comprehensively performed? In many respects, it appears this has yet to take place, as we did not identify any documented requirements for research data or digital collections preservation that could be leveraged to build out technical, programmatic, or policy capabilities.

In many ways the lack of detail on digital preservation is not surprising. Large-scale digitization is exciting. The prospect of undertaking this at unprecedented levels in order to enable broad access to some of the world's most diverse collections and research output is motivating. Digital preservation in and of itself is a rather underwhelming topic by comparison, so it has been somewhat buried here and overshadowed by the alluring prospects that seemingly result from digitization alone. Yet digital preservation is an essential underlying component if that digitization investment is to live up to its promise, especially over time. As a point of comparison, building a new museum, such as the recently opened National Museum of African American History and Culture is equally exciting. But you don't undertake an architectural project without first ensuring that the land is prepared to support the structure, and that the foundation is laid. You don't put up walls until you have the frame, just as you don't put collections on exhibit until you install appropriate lighting, HVAC, and ensure security. And of course, once the building is complete, you allocate skilled staff and an adequate level of finances to its maintenance.

A successful and sustainable set of digital resources requires a similar plan. Lay the foundation (e.g., storage), put up the frames (e.g., management databases and software), ensure security (e.g., geographic redundancy, security risk management, etc.), and maintain (e.g., skilled staff, monitoring, preservation intervention).

[36] Smithsonian Institution. Creating a digital Smithsonian: Digitization strategic plan, 2010-2015. p. 11. https://www.si.edu/content/pdf/about/2010_SI_Digitization_Plan.pdf. Accessed September 26, 2016.

4.2 SD 600 & SD 610

SD 610, an important policy outcome of the 2010-2015 Digitization Strategic Plan, lists several advantages that effective resource management will enable, including increase in quality, usability, and access to digital assets, as well as the preservation of, "digital assets that are in danger of loss due to deterioration or obsolescence,"[37] amongst other goals for access and usability. Digital asset is defined in this document as:

> Content that is recorded and transferred in a digital format. It may include text, still images, moving images and sound recordings, **collections that are digital** (i.e., digital art), **research datasets and other types of media originally created in digital format or digitized from another format or state** (i.e. a digital surrogate) that are created, stored, or maintained by the Smithsonian. For the purpose of this directive, digital assets also include metadata used to describe the digital asset and its content.[38]

Notably, this definition is not specific to collections, as it includes research data, and other types of digital institutional output. It also emphasizes that assets aren't just the product of digitization, they may be born-digital as well.

This definition, and others found in SD 610 offer clarity on aspects of digital resource management that are absent, or outdated in SD 600. Definitions for terms such as life-cycle management, metadata, digital preservation, and trusted digital repository (TDR) provide usage guidelines for commonly used and contemporary terms. By comparison, SD 600 still reflects the parlance of the time it was written (2001), which is difficult to map to the newer terminology of SD 610. Because SD 600 hasn't been updated in quite some time, and because there has been no effort to sync newer SDs like SD 610 to this one, how these digital terms should provide guidance on digital collections policies, which should be created per the mandate of SD 600, remains unclear. Differing interpretations of SD 600 for digital collections was identified as a problematic issue by interviewees, presenting risks to the longevity of these resources.

Returning to SD 610, despite mention in the background and scope that the directive applies to digital assets in a broad sense, including those created by research centers, offices, and programs, and that it addresses the full asset life cycle, in actuality it provides very limited direction beyond collections digitization. In fact, it leaves a great deal of ambiguity when it comes to 1) born-digital assets, and 2) resources other than collections, such as research and other institutional output.

[37] Smithsonian Institution. Smithsonian Directive 610: Digitization and digital asset management policy, March 31, 2011. p. 2. https://www.si.edu/content/pdf/about/sd/SD 610.pdf. Accessed September 26, 2016.
[38] Ibid., 4.

Like other Smithsonian Directives, SD 610 contains sections addressing principles, background, scope, definitions (like the one above), and roles and responsibilities. Several gaps in these areas are noticeable:

- The scope states that the directive, "applies to all units that acquire, create, or maintain Smithsonian digital assets," which as noted above includes research datasets and other types of media. Yet in the next sentence, it further specifies that the directive, "covers collections that are digital, which are also subject to SD 600."[39] While it could be interpreted that the directive pertains to collections *in addition* to other types of assets, the special emphasis placed on digital collections seems to limit the applicability of this document to collections only. Other resource types are not discussed with any further specificity.

- Roles and responsibilities are largely limited to those for digitization, particularly when looking at the responsibilities of the Secretary, Under Secretaries, OCIO, and National Collections Program. There are no roles defined for stewardship of digital assets beyond the role of units and the DPO. There is also no group assigned responsibility for research data or institutional output, although the National Collections Program is explicitly assigned the task of supporting collections, furthering the sense that this directive in fact does functionally apply to those other resource types.

- Little responsibility for life-cycle management is actually mandated. The Digitization Program Office (DPO) is, "responsible for improving the overall stewardship and long-term management of the Smithsonian's digital assets by providing leadership and policy oversight of the pan-Institutional digitization program." The DPO also, "assists units in developing project digital asset management plans."[40] In fact, DPO is the only group with the role of advising units on long-term management, and yet they are explicitly tasked with doing this through digitization leadership, not preservation guidance. While this is logical — they are the *Digitization* Program Office, after all — this leaves nowhere for units to turn for long-term preservation support.

In conclusion, despite this document highlighting digital preservation as an essential function, and providing some good initial definitions that pertain to digital stewardship, no detail is actually provided on how preservation should be achieved. Aside from identifying that it is the unit's responsibility to develop digital asset management plans, which, "must cover the full data life cycle (from planning for data creation to accessible archiving, preservation, and possible disposition) for each unit project that creates or collects digital assets,"[41] no other roles and responsibilities for preservation are described. The directive leaves the responsibility of preservation solely in the hands of the units, without providing any guidance, or real mechanism

[39] Smithsonian Institution. Smithsonian Directive 610: Digitization and digital asset management policy, March 31, 2011. p. 3. https://www.si.edu/content/pdf/about/sd/SD 610.pdf. Accessed September 26, 2016.
[40] Ibid., 7.
[41] Ibid., 9.

for measuring accountability. It falls short of fulfilling the need for a digital preservation policy, guidance on implementation at the unit level, or the specifying role that central services should play, all of which by contrast are defined with regard to digitization. Additionally, while the directive states that it is applicable to, "all units that acquire, create, or maintain Smithsonian digital assets, prepare them for data interchange and interoperability to support sharing and repurposing, or manage other life-cycle functions of these assets,"[42] which includes research data and other output, it does not provide any guidance on how this data should be stewarded.

4.3 Digital Asset Management Plans / Data Management Plans

SD 610 prescribes that, "units shall develop digital asset management plans for every unit project that collects or creates digital assets."[43] The Smithsonian Digital Asset Management Plan template includes a section on life-cycle management, which states: "The Life Cycle Management section clarifies how the unit stewards its digital assets to ensure these assets are appropriately created and cared for through their intended lifespan. Proper stewardship ensures that digital assets will not be 'orphaned' or compromised in a way that results in data loss."[44] This is the primary mechanism through which collecting units define their strategy for preservation of the digital assets that result from a digitization project.[45]

Although DAMPs are primarily used by collecting units, they are quite similar in function and content to the data management plans (DMPs) that are increasingly required by funding agencies for research initiatives, including many federally-funded grants received by Smithsonian researchers. In the United States, the mandate for data management comes from a 2013 memo from the the White House Office of Science and Technology Policy (OSTP), which specifies that, "all extramural researchers receiving Federal grants and contracts for scientific research and intramural researchers develop data management plans, as appropriate, describing how they will provide for long-term preservation of, and access to, scientific data in digital formats resulting from federally funded research, or explaining why long-term preservation and access cannot be justified."[46] SI DAMPs and data management plans are very similar in content and structure, as can be seen in Table 20.

---

[42] Smithsonian Institution. Smithsonian Directive 610: Digitization and digital asset management policy, March 31, 2011. p. 3. https://www.si.edu/content/pdf/about/sd/SD 610.pdf. Accessed September 26, 2016.
[43] Ibid., 9.
[44] Smithsonian Institution. Digital Asset Management Plan template, January 2013. https://www.idigbio.org/wiki/images/2/20/NMNH_Digital_Asset_Plan_Template.pdf. Accessed September 28, 2016.
[45] As noted above, these plans are primarily being created by collecting units to describe their plans for the result of digitization project. Also the definition for digitization in this document, "A set of processes that converts physical resources to a digital form, or that creates materials in a digital form (born digital)," is out of sync with the common usage of the term today, which would exclude born-digital.
[46] Holdren, John P. Increasing access to the results of federally funded scientific research. Executive Office of the President, Office of Science and Technology Policy, February 22, 2013. https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf. Accessed September 26, 2016.

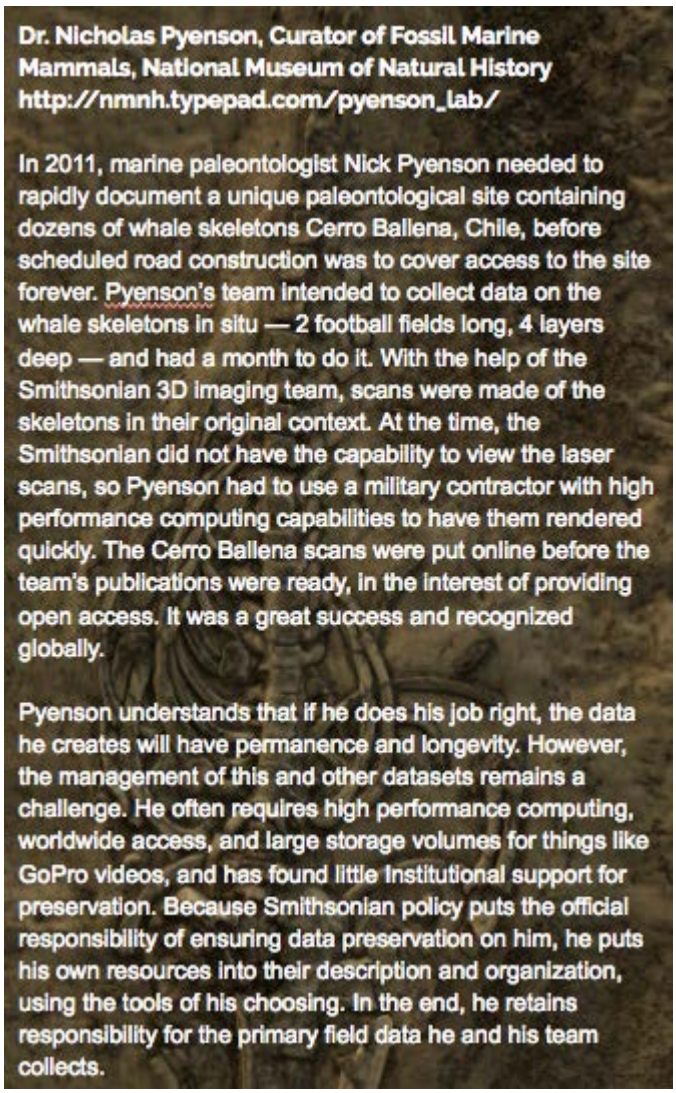
SI Digital Asset Management Plan Requirements[47]	NSF Data Management Plan Requirements[48]
• Categories and volume of assets, asset types to be produced • Metadata standards, structures, and values to be used • Asset usage goals, audiences, and interoperability expectations • Policies for access, attribution, and restrictions • Lifecycle management plans, including designated steward of the digital assets, data storage environments and physical locations, disaster recovery plan, plan for securing sensitive or personally identifiable information, risk assessment, intended lifespan, and reporting requirements • Short- and long-term storage requirements, non-centrally supported hardware and software	• Types of data to be produced during the project • Standards to be used for content and metadata format, or proposal for format • Roles and responsibilities for management[49] • Policies for access and sharing, including provisions for privacy, confidentiality, security, and intellectual property • Policies and provisions for re-use, re-distribution, and production of derivatives • Plans for archiving and preservation of data

Table 20. Comparison of SI digital asset management plan and NSF data management plan requirements

It is logical that these plans are required. Researchers and curators are asking for funds to create what are purported to be valuable digital outcomes; it should be expected that they take responsibility for the final output to ensure accessibility of these publicly funded resources over the long-term. Funders want stakeholders to be aware of what it takes to steward digital resources, and requiring these plans asks that they think through what it will take to manage those assets over time.

[47] Smithsonian Institution. Sample digital asset management plans. Supplied June 7, 2016.
[48] *For example data management plan requirements see:* National Science Foundation. NSF 15-1: Chapter II - Proposal Preparation Instructions, December 26, 2014. http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp#dmp. Accessed September 26, 2016.
[49] *For data management plan requirements specific to the Biological Sciences Directorate see:* National Science Foundation. NSF 15-1: Chapter I - Pre-Submission Information, December 26, 2014. http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_1.jsp#IG2. Accessed September 26, 2016.

The problem with these approaches, particularly from the Smithsonian's perspective as an investor and stakeholder, is that the mandates are not matched with any Institutional commitment to preservation and access. There are no guidelines, no recommendations or stated preferences on how data should be stored, no listing of acceptable repositories, no guidance on metadata. While this sort of information doesn't necessarily make sense to include in the guidelines for a DAMP or DMP because the details will change over time, they should still be made available.



We see three core problems with the DAMP or DMP model with regard to digital preservation:

1. They put the responsibility for digital preservation entirely on the content creator.

2. They are project-centric, rather than Institutional or even unit-centric.

3. There is little to no accountability for digital assets at the Institutional-level.

Without a program in place to provide necessary support to content creators who likely have no expertise in digital preservation, grant funds and Institutional resources are used inefficiently, and digital assets are placed at risk of loss once the content creator moves on to the next project. Today there is no group at the Smithsonian with the role of ensuring that there is follow through on the preservation aspects of these plans that keep the researcher's and Institution's interests in mind, or that the project's requirements are matched to the Institution's capabilities.

4.4 Sharing Smithsonian Digital Scientific Research Data from Biology

In March 2011, the Smithsonian Office of Policy and Analysis issued a report titled, "Sharing Smithsonian Digital Scientific Research Data from Biology." Recognizing that the 2010-2015

Smithsonian Strategic Plan made accessibility to digital scientific resources a priority, the report set out to provide, "an overview of the issues, challenges, and opportunities that the Smithsonian and the wider scientific community face as they work to increase access to and use of the growing volume of digital data produced by the world's researchers."[50] Although this report pre-dates the White House OSTP memo, it anticipated the growing need to share research data. It focuses on the biological sciences, though it may be extended to other scientific domains.

The authors reach several conclusions relative to the question of life-cycle management and preservation of research data:

1. The biological sciences largely adhere, "to a traditional small-science approach to data management in which individual research teams see their data as proprietary and pay little attention to the data management necessary to facilitate their use by others or their long-term preservation."[51] The disconnected datasets that result are difficult to discover, access, and use, and many are at great risk of loss.

2. The Smithsonian lacks an Institutional strategy to guide progress, and promote systematic data management and sharing, and instead approaches are, "small-science, seat-of-the-pants, fragmented, and mostly unit- or department-based,"[52] and furthermore, that the, "absence of an overarching Institutional strategy and framework for data management and sharing has contributed to fragmented and often opportunistic efforts in this area."[53]

3. Researchers are not dependable data stewards as they, "Currently have virtually no incentives, and many disincentives, for engaging in data management beyond the minimum required for their own analytic purposes."[54]

The authors' recommendations for resolving these challenges were broad and holistic, addressing the problems with a variety of approaches designed to mitigate loss, distribute the share of responsibilities, and ensure that the appropriate human, technical, and financial resources could be efficiently leveraged. These include, amongst others:

- Provision of a status for Smithsonian digital biology data comparable to that of the National Collections covered in SD 600;

[50] Smithsonian Institution. Sharing Smithsonian digital scientific research data from biology. Office of Policy and Analysis Research Reports, 2011. p. v. https://repository.si.edu/handle/10088/26386. Accessed September 26, 2016.
[51] Smithsonian Institution. Sharing Smithsonian digital scientific research data from biology. Office of Policy and Analysis Research Reports, 2011. p. xi. https://repository.si.edu/handle/10088/26386. Accessed September 26, 2016.
[52] Ibid., xiii.
[53] Ibid., xvi.
[54] Ibid., xiv.

- Development of core requirements for digital biology data management over their life cycle, including data management standards, specifications for metadata, and acceptable formats;
- Infrastructure to support data management and sharing, including a trustworthy digital repository for long-term preservation;
- Criteria for determining the appropriate level of data management for specific data sets;
- And design of an organizational structure to support data management and sharing at all levels, including the definition of roles and responsibilities for Smithsonian central support offices (particularly OUSS, OCIO, SIL, and SIA) and research units.

The findings of this previous study are consistent with ours. During the discovery process for the current study, we heard of similar challenges expressed and similar recommendations suggested. In fact, we found that very little has changed in the past 5+ years since this report was written, and the report itself seems to have at little lasting impact; it was largely unfamiliar to the members of the Digital Preservation Working Group. The problems the authors highlighted still persist, perhaps have become more entrenched, and little effort has been made toward implementing their recommendations. We agree that the need for such changes is critical, and support these recommendations today.

By comparison, a similar study by the Office of Policy and Analysis, “Concern at the Core”[55] (2005), which looks broadly at collections management, seems to have had measurable impact. Most, if not all of the recommendations of that report, appear to have been systematically tackled, if incrementally, over the years, and it was pointed to by several interviewees as a great example of a foundational document for pan-Institutional initiatives. Its success may in part be attributed to a recommendation it makes that the National Collections Program (NCP) be given the authority to as a central advocate and policy office for collections. As a result, NCP was then able to work with units to carry out additional recommendations in the report, to provide oversight, and act as a liaison between collecting units and Smithsonian Senior Leadership. It appears that a similar role has not been put in place for research data management. It may be useful to re-evaluate these two studies to identify why one made more of an impression than the other.

[55] Smithsonian Institution. Concern at the core: Managing Smithsonian collections. Office of Policy and Analysis Research Reports, April 2005. https://www.si.edu/content/opanda/docs/Rpts2005/05.04.ConcernAtTheCore.Contents.pdf. Accessed September 28, 2016.

5 SUMMARY OF FINDINGS

The interviews, survey, and documentation review have lead us to reach several conclusions, the most central being that digital preservation is not optional for the Smithsonian; it is a core function, essential to the Institution's mission. Some excellent work in this area is being conducted by members of the DPWG, and other stakeholders, who work tirelessly to ensure the digital resources in their care are safeguarded. The mature and robust enterprise DAMS, largely compliant with digital preservation standards, provides an essential central preservation service to a large number of Institutional digital assets. An acknowledgement by Senior Leadership that digital preservation is an essential part of the organization's mission, and that the current state needs improving, helps elevate the importance of the needs in this area.

Yet despite these and other strengths, digital preservation is not occurring systematically today. A close examination at the organizational infrastructure, technological infrastructure, and resources framework across the Institution through the lens of the interviews, researcher survey, and document review, has revealed that this can be attributed to several factors, which are described below.

5.1 Conclusions

Existing policies that address digital resources are very ambiguous with regard to digital preservation. While they provide a great deal of detail on digitization, and more specifically, digitization of collections, they lack specificity on the responsibilities and tasks required to support the digital assets that result from the digitization process over time. Furthermore, these policies provide almost no guidance on born-digital resources or on non-collections content such as institutional output (e.g., event recordings) or research data. *See Sections 4.2 and 4.3.*

Existing policies put preservation responsibility solely in the hands of the resource creator. SD 610 specifically dictates that the department or individual responsible for the creation of digital resources (primarily through digitization, but also implicitly through processes that generate born-digital resources) is responsible for their preservation. There are no further roles for preservation defined in this document.[56] *See Section 4.2.*

The Smithsonian does not offer formal preservation support to creators or stewards. There is no official central office to provide guidance or oversight, no explicitly dedicated infrastructure, no guidelines, no policies, no staff tasked with preservation management support. The resources that do exist, such as the DAMS, play an *ad hoc* preservation role, as they have not been officially charged with providing this service. *See Sections 2.2.1, 2.3, and 4.2.*

[56] It has been noted by several members of the DPWG, who helped write SD 610, that when this document was issued, the focus was on digitization and publication of digital assets, and only marginally on digital preservation. Putting the responsibility on units and content creators made sense at the time, however, the group agrees today that this was insufficient, and that central roles are also necessary.

Content creators and collecting units clearly state that performing preservation in addition to primary responsibilities is an unrealistic expectation. Furthermore, they also don't necessarily have the expertise to ensure that digital resources will remain accessible over the long-term. Leaving valuable digital assets in the care of those who are busy and unskilled in preservation management places those resources at risk of loss, and contributes to the accumulating backlog of digital resources that are largely inaccessible today. *See Section 2.3.*

Because supporting roles of digital preservation are not defined, creators of digital resources unsure where to turn for help. In lieu of a formal digital preservation program, content creators and stewards look to a myriad of stakeholders for help: unit IT heads, Smithsonian Institution Libraries or Archives, SI DAMS, or collaborating partner Institutions such as a university. Because there are no official policies for long-term data stewardship, each of these groups may provide very different recommendations and support. For many, the challenge of entrusting digital resources to a third-party is daunting, complex, burdensome, or even risk prone. For these stakeholders, storing data on a hard drive in their office is the *de facto* response. As a result, there are large volumes of data sitting on unmanaged storage. *See Sections 2.3 and 3.*

The project-centric approach of data management right now is inefficient, wasteful, and places digital assets at risk of loss. It is easy to store the files that result from a research project or small digitization effort on a set of hard drives or in Dropbox. It is harder to manage those files when, over the years, thousands of projects accumulate on such media, and neglect to create metadata to make those files findable and understandable again. Staff at SI DAMS, which often becomes the repository for project output, report enormous challenges when the digital resources are eventually submitted, such as corrupt data, obsolete formats, and missing metadata. Furthermore, because project funding is temporary, there are no funds to allocate toward more reliable storage, or make corrections or improvements to data after the fact. The problems noted by SI DAMS likely arise when project funding is over. *See Sections 2.3 and 4.3.*

The volume of potential digital assets that require long-term management is enormous. Because digital preservation responsibility is designated on a per-project basis, there is no easy way to assemble a global view of the digital resource responsibilities of an individual, department, or unit, much less the entire Institution. However, according to data gathering throughout this study, we estimate that there are:

- **6.7 PB of research data** currently scattered across the Institution, third-party storage services, and partner Institutions.[57] 86% of this total is unmanaged, and 35% is not currently under SI control. By 2026, the total is estimated to be 280 PB. Researchers report that 60% of these resources should be held indefinitely. *See Section 3.1 and 3.4.*

[57] It was noted by the DPWG that this number in fact might be low, however the group feels that it is an appropriate starting point for conversation.

- Nearly **2 PB of collections and institutional output** with potential value if one counts the holdings of DAMS, SIA, local network drives at units, assets stored OCIO-maintained central Isilon storage, and third-party services such as the Internet Archive. Only a small percentage of these digital resources are truly managed as assets, meaning they have the metadata to make them actionable over time. The rest holds potential, but that slips away daily. This number is also likely to increase dramatically in the near future once rapid capture efforts are accelerated. *See Section 2.2.*

Lack of a mechanism to quantify the scale of existing digital resources and pace of growth has been paralyzing. Interviewees report that it is difficult to address the global digital preservation problem when to date it has been unclear what exactly that means. Depending on the stakeholder's perspective, they may feel that research data is a drop in the bucket when compared to collections digitization output. Others feel the reality must be quite the opposite. Indeed, according to our estimations, which are extrapolations based on a sample, the reality may very well be that there is a particularly large volume of research data potentially of long-term value across the Institution that is currently greater than the total holdings of digital collections. *See Section 3.1.*

Lack of clarity around definitions of digital resource types seems to inhibit policy creation and designation of responsibility. Interviewees expressed confusion over the divisions between what is SD 600 digital collections, what is associated information, and what is Institutional output (e.g., whether this category includes art conservation research output, collections information, etc.). Further confusion was expressed over what exactly constitutes "research data," and whether this destination only includes Derived Data or also Primary / Raw Data gathered in the field. On the collections side, part of the problem seems to be that SD 600 is outdated with regard to digital information, and therefore it becomes difficult for units to interpret for local policies addressing digital resources. On the research data side, the lack of any policy that would designate value seems to be particularly inhibiting. In short, it is unclear exactly what should be the target of preservation. *See Sections 2.3.1, 2.4 and 4.2.*

There are functional and storage gaps in the existing infrastructure, if all digital resources of value are to be managed by the Institution over the long-term. The combined capacity of existing digital repositories is no where near the potential volume of assets. Furthermore, existing service offerings of these repositories leaves many types of resources without a logical home. For example, SIL's DSpace Digital Repository accepts submissions of data sets that accompany publications, but not Primary or Derived data. SIdora, which is not in full production, accepts datasets of any type or format, but doesn't commit to long-term preservation, only management for active research. SI DAMS currently accepts images, video, and audio, as well as time-based art, but not textual, document, or other formats that are common to web, email, design, and other archives. As a result, there are large volumes that lack any management whatsoever. *See Section 2.2.*

Without designating shared responsibilities for all aspects of digital preservation, complacency persists. At the moment, no individual or group is tasked with oversight and

accountability for digital preservation. And there is certainly no one with the authority to take decisive action to improve the current state. There are several stakeholders with digital preservation expertise (i.e. the DPWG), but they are tasked with looking at their unit or service, not with the whole Institution. These stakeholders are mostly in the collections arena too, not in research, leaving researchers without a unified voice or representation.

This situation results in silos of effort, such as the SI DAMS, SIA Digital Archives, SIdora, and grassroots efforts such as the pan-Institutional Time-Based Media and Digital Art Working Group.[58] Excellent groundwork has been laid by these dedicated teams, and the relative security of digital collections can largely be attributed to their efforts. But there is a clear lack coordination and oversight. As a result, incredibly large volumes of digital resources are falling through the cracks, unidentified, unmanaged, and as time passes, face the very real prospect of loss. *See Sections 2.3, 2.4, 3, 4.2, and 4.3.*

5.2 Organizational Maturity

A long-standing starting point for determining the maturity of an organization's digital preservation program is Anne R. Kenney and Nancy McGovern's 2003 paper, "The Five Organizational Stages of Digital Preservation." This paper identifies five stages of organizational response to digital preservation that emerge as a result of increased experience. These are:

1. **Acknowledge**: Understand that digital preservation is a local concern;
2. **Act**: Initiate digital preservation projects;
3. **Consolidate**: Segue from projects to programs;
4. **Institutionalize**: Incorporate the larger environment; and
5. **Externalize**: Embrace inter-Institutional collaboration and dependency.[59]

Their simple definition and description of each level provides a clear method for institutions to identify what stage they are in, and can be used as an ongoing reference as digital preservation programs progress, ultimately looking toward the criteria in ISO 16363 -- Audit and Certification of Trustworthy Repositories[60] as comprehensive digital preservation benchmarks.

Considering the Smithsonian Institution's current state against Kenney and McGovern's readiness criteria, we find that the organization is moving toward stage 2 (Act) for research data, and stage 4 (Institutionalize) for collections, given that the research domain is dominated by project-centric approaches, but that there are maturing workflows and technologies in the collections arena. Key indicators for each stage are summarized in Table 21.

[58] Smithsonian Institution. TBMA: Time based media art at the Smithsonian. https://www.si.edu/tbma. Accessed September 26, 2016.
[59] Kenney, Anne R. and Nancy Y. McGovern, "The five organizational stages of digital preservation," *Digital libraries: A vision for the 21st century,* 2003. Ann Arbor, MI: Michigan Publishing, University of Michigan Library. http://quod.lib.umich.edu/s/spobooks/bbv9812.0001.001/1:11/--digital-libraries-a-vision-for-the-21st-century?rgn=div1;view=fulltext. Accessed September 26, 2016.
[60] See the public version at: https://public.ccsds.org/pubs/652x0m1.pdf

Stage 2 - Key Indicators	Stage 4 - Key Indicators
• **Policy and planning:** the preservation policy may remain implicit in stage 2 or may be expressed in general terms, though evidence that the organization is committing to digital preservation accumulates. • **Technological infrastructure:** the organization may stipulate a set of technical requirements that apply to each project, or, more likely, will devise technical requirements that are project-specific and reactive. Digital content may be dispersed across multiple servers in multiple locations or be co-located using available equipment, depending on the size of the projects, the level of technology support obtained for the project, and the nature of technology support within the organization. Cross-project technology planning is less likely to occur at stage 2 than at later stages. • **Content and use:** efforts may go deep into addressing the range of requirements for selected types of digital materials or collections, or address some or all collections in basic ways.	• **Policy and planning:** organization-wide entities that coordinate, authorize, and mandate digital preservation programs may be established, or some equivalent mechanism that allows for consistent and systematic management rather than event-based responses; establishing a comprehensive policy framework provides the focus for planning efforts. The framework as outlined and populated will address, in some way, all components of ISO 16363. • **Technological infrastructure:** beginning of true technology planning and management, characterized by responding to rather than reacting to, and anticipating needs; the infrastructure may be distributed rather than centralized, but investments in infrastructure are more likely to be based upon requirements that are defined and approved at a high level of management, and implemented across the organization. • **Content and use:** rather than presuming that all digital materials will be preserved as part of the organization's commitment to digital preservation, the implications of that commitment are more fully understood and acceptance criteria is established and utilized to determine the scope of collections that will be actively preserved by the organization. Services to capture, store, maintain, and provide access to digital resources become integral to the organization and subject to relevant monitoring and measurements, and expectations that these services will be reliable and consistent become evident.

Table 21. Summary of key indicators for stages 2 and 4 from "The five organizational stages of digital preservation." Our conclusions are that research data is in stage 2, but collections are moving toward stage 4.

These indicators resonate soundly with our findings.

5.3 Threats

The Simple Property-Oriented Threat (SPOT) Model for Risk Assessment[61] defines six essential properties that digital objects exhibit if preservation efforts are successful:

- **Availability**: A digital object is available for long term use by having been ingested into and maintained in a preservation environment;
- **Identity**: A digital object can be referenced, discovered, and retrieved;
- **Persistence**: The bits that make up a digital object remain available and uncorrupted;
- **Renderability**: A digital object can be used in a way that maintains its significant properties or characteristics;
- **Understandability**: All associated information needed to guarantee a digital object can be interpreted and understood by users; and
- **Authenticity**: A digital object is what is purports to be.

The SPOT Model can be used to examine the maturity of a preservation and access environment according to its data management and preservation practices and how well these mitigate threats to content longevity. By considering the state of the Institution's assets against the six essential properties of preserved objects defined by the model, we can gauge which threats are present and require mitigation. Based on these criteria, we feel that the threats outlined in Table 22 should be addresssed.

[61] Vermaaten, Sally, Brian Lavoie, and Priscilla Caplan. Identifying threats to successful digital preservation: The SPOT model for risk assessment. *D-Lib Magazine*, September/October 2012, Volume 18, No 9/10. http://www.dlib.org/dlib/september12/vermaaten/09vermaaten.html. Accessed September 26, 2016.

Property	Threat	At Risk	Notes
Availability	A digital object is not selected for preservation, either intentionally or unintentionally, and subsequently disappears	Research data	Only 13.2% of research data is in a managed environment today
Identity	Sufficient metadata is not captured or maintained.	All SI digital assets	This threat is present for a large amount of research data, as well as some digital collections items, whose metadata and other associated information required for long-term accessibility is not always provided the same level of care as the file objects, or may not exist at all
Persistence	Useful life of storage medium is exceeded (e.g., media obsolescence, mean time to failure exceeded)	All SI digital assets	All digital assets stored on hard drives, optical media, and unmanaged offline media are at risk
Renderability	The appropriate rendering environment (hardware and software) is unknown (e.g., the format of the source object is unidentifiable)	Research data	If not documented, the rendering environment required to reproduce research data is likely to be unknown
Understandability	The entire representation network is not obtained or archived, with the consequence that supplementary information is itself not understandable in the future.	Research data	Capturing just the primary bits is insufficient if additional context required for reproducibility

Table 22. Relevant SPOT threats

Many of these threats are particularly salient to reproducibility of research data, which is a primary incentive for their preservation, and rationale behind the mandate to create research data management plans. A recent report that was the outcome of a National Science Foundation Directorate of Mathematical and Physical Science workshop with research communities, recommends that, “Data upon which publications are based should be available in machine-readable digital format, and persistently linked to those publications.” They further specify additional aspects of a research project that should be stored along with the data outcomes, where relevant. These include:

- Software: the software used to create, process, and analyze the data

- Workflow: instructions, frameworks, or scripts used to run the software
- Software environment: a specification or an instantiation of the requisite operating system, architecture, libraries, machine state, etc., that are necessary to run the software/workflows
- Simulation capabilities: the capability to run the software with different parameters than used to generate the original data
- Documentation: a description of the software, workflows, and other information describing how the data were derived, processed, and analyzed
- Data characterization: documentation of data (formats, content, etc.), and the metadata that describes it and makes it discoverable and re-useable

If these components are gathered and packaged with the data outputs, preserved, made accessible, and are ensured to be reusable, the data meets the gold standard of reproducibility, and is much more likely to contribute to the advancement of human knowledge. The pyramid in the Figure 11 illustrates the increasing value of research data as it is receives additional levels of curation. Preservation of data is really just one step above the bare minimum, however, if the Smithsonian wants the data generated under its umbrella to have the potential to reach the top of the pyramid, it is important to start here.

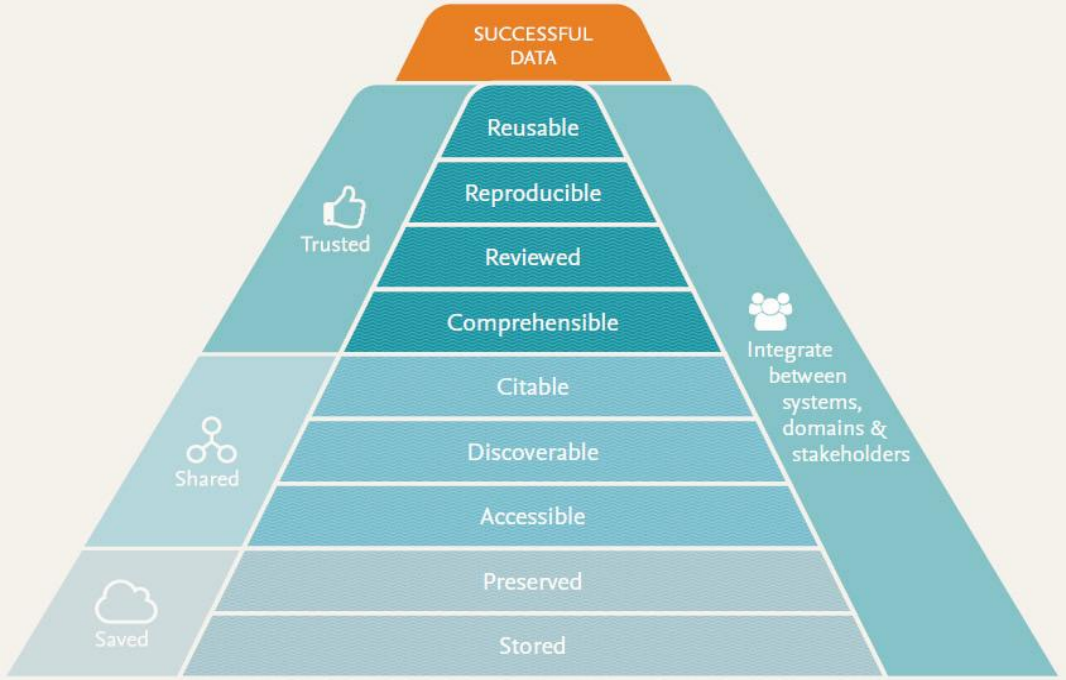


Figure 11. Illustrates the hierarchy of research data as it receives increasing levels of curation, from "10 Aspects of Highly Effective Research Data"[62]

[62] de Waard, Anita, Cousijn, Helena, and Aalbersberg, IJsbrand Jan. 10 Habits of Highly Effective Research Data. Elsevier Connect, December 11, 2015. https://www.elsevier.com/connect/10-aspects-of-highly-effective-research-data/_nocache

6 RECOMMENDATIONS

Shaping the future by preserving our heritage, discovering new knowledge, and sharing our resources with the world.

— Smithsonian Institution Vision[63]

Through investment toward the creation of a Digital Smithsonian, the Institution recognizes that the future it aims to shape is a digital one. This investment amounts to millions of dollars that has gone toward digitization, acquisition of digital collections, and research. Preserving that investment is not optional, it is an essential function of the organization. Long-term maintenance of research data enables future research; stewardship of digital collections items enables long-term, global access.

The greatest challenge to enacting an effective digital preservation program today is likely that the digital reality is still a new one. Since its establishment in 1846, the Smithsonian has been responsible for preserving some of the richest, most varied, and most significant museum, library, and archive collections in the world. It isn't as if the preservation of these materials is solved; improvements to physical collections preservation are ongoing today. However, the Institution is now tasked with the challenge of continuing to maintain its physical resources, while adding digital preservation to its responsibilities. Implementing systematic digital preservation will require new organizational structures, policies, resource allocation, technologies, and responsibilities.

The final report of the National Science Foundation-sponsored Blue Ribbon Task Force on Sustainable Preservation and Access outlined five economic conditions required for sustainability of digital resources:

- Recognition of the benefits of preservation by decision makers;
- A process for selecting digital materials with long-term value;
- Incentives for decision makers to preserve in the public interest;
- Appropriate organization and governance of digital preservation activities; and
- Mechanisms to secure an ongoing, efficient allocation of resources to digital preservation activities.[64]

[63] Our Vision. Smithsonian Institution. http://www.si.edu/about/mission. Accessed September 28, 2016.
[64] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Final report of the Blue Ribbon Task Force on sustainable digital preservation and access, February 2010. p.12. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf. Accessed September 28, 2016.

The recommendations provided below are intended to be the starting point for the creation of a digital preservation program that is aligned with these conditions. It is our hope that by laying a strong foundation, the Smithsonian will be in a position to ensure that the opportunity for a future scholar to learn from and use these rich resources in the creation of her own discoveries will never be lost.

6.1 Instill a sense of urgency

Experts agree that successful change efforts start by galvanizing stakeholders with a sense of urgency about the need to do something.[65] The Secretary has recognized the current digital preservation gap, and has assembled a working group to begin to address it. These are important components of initiating change, however, even the subject matter experts in the Digital Preservation Working Group — who acknowledge the need to act — have trouble broadly communicating the urgency with which the problem needs to be addressed. This is in part because the challenge has not been quantified, making it difficult to tell the story of just how big the unmanaged data problem is. Furthermore, the differences in data formats, means of production, and current storage efforts for research data and collections are so different now, it is difficult to think about bringing both into a single, overarching digital preservation program.

6.1.1 Quantify the need

As noted in Section 3.1, the results of the survey show that the total estimated volume of research data that can be found across the Smithsonian amounts to approximately 6.7 PB. An important insight gained in looking at survey responses is the amount of data not managed in an environment governed by preservation policies and practices, which puts them at significantly increased chance of loss. Extrapolating from the survey results, we estimate that 86% of total reported research data holdings, or potentially 5.8 PB is unmanaged. This is a preliminary calculation, but it provides a handhold to quantify the challenge and start the discussion. The important, and concerning, takeaway is that the vast majority of this data is stored on local storage media, such as hard disk drives (HDDs), optical media, USB drives, and small NAS devices. Furthermore 60% of the total research data holdings (6.7 PB), were reported by researchers to have permanent retention value. It can be extrapolated that a large percentage of those unmanaged data are also of long-term value.

To put the total volume of potentially unmanaged data in perspective, that 5.8 PB would be the equivalent of 1,933 three TB desktop hard drives. In reality, it is not hard to imagine that number of drives across the Institution. And, in fact, 6.7 PB itself is not such an alarming number when compared to other medium-to-large organizations, many of which are expanding their data centers out to multi-petabyte scales. The Library of Congress, for instance, currently has a

[65] This need has been popularized by John P. Kotter, one of the predominant voices in change management today. Kotter's 8 step process for leading change is the subject of numerous publications, including *Leading Change* (first published 1996, revised 2012). "Create a sense of urgency" is Kotter's first step. "Leading Change: Why Transformation Efforts Fail" is *Harvard Business Review*'s best seller on the topic of change management.

preservation storage capacity of 166 PB, and they grew by 2.9 PB in 2015 alone. Recent studies of higher education Institutions across the UK found similar results to ours: an average of 5 TB per researcher[66] (researchers at the Smithsonian reported 6.7 TB on average), the bulk of which is on unmanaged storage.[67] Total research data volumes across institutions range from approximately 14 PB at the University of Oxford at the high end, to 4.2 PB at the University of York at the low end.[68] The author of the latter study emphasizes that, "Although research income might be flat, data volumes are rising, and expected to rise. This is due to the falling cost of creating data,"[69] noting that some today's consumer devices can produce exponentially more data than in the past: a GoPro high definition video camera left running for an extended period can result in TB of high definition video.

What is alarming about the volume at the Smithsonian is not the size, but the fact that this much data (and, potentially more) is distributed across the Institution on media that are extremely vulnerable: likely not backed up, not networked, not accessible, not documented or described, and whose contents are likely only known to their creator. The scale of this unmanaged data in the research domain is of tremendous cause for concern.

The situation is not nearly as dire on the collections side; most units submit their digital assets to the DAMS, which takes on a *de facto* preservation role. However, the anticipated growth in digital resources as an outcome of ongoing digitization, rapid capture, 3D scanning, and born-digital collections accessioning necessitates a closer look at requirements for storage, staffing, procedure, policy, and preservation functionality that will be necessary to support these resources over the long-term. And in reality, the picture may be more alarming than it appears at the moment. For example, it has been reported that many digital collections objects are kept on removable media, not backed up, and have been either lost or corrupted over time. These digital files need to be quantified to get a clear view of the collections landscape.

Finally, regardless of whether the digital files themselves are stored in a managed preservation environment, the existence of associated metadata remains in question. A 2014 IDC Digital Universe study on the growth of data around the world found that "from 2013 to 2020, the digital universe will grow by a factor of 10 – from 4.4 trillion gigabytes to 44 trillion. It more than doubles every two years."[70] They note, however, that the majority of this data is not truly useful. To achieve its potential value, data should be characterized and/or tagged with metadata. In 2013, only 22% of the entire digital universe (4.4 trillion gigabytes) fit into this category. By 2020, they estimate this number could reach 37%.

---

[66] Addis, Matthew. Estimating Research Data Volumes in UK HEI. figshare, 2015, 5. https://dx.doi.org/10.6084/m9.figshare.1575831.v5.
[67] See, for example, Parsons, Thomas; Grimshaw, Shirley; and Williamson, Lurian. Research Data Management Survey. The University of Nottingham, June 2, 2013.
[68] Addis, Matthew. Estimating Research Data Volumes in UK HEI. figshare, 2015, 5. https://dx.doi.org/10.6084/m9.figshare.1575831.v5.
[69] Ibid., 9.
[70] EMC and IDC. The digital universe of opportunities: Rich data and the increasing value of the Internet of Things — Executive summary. EMC2, April 2014. http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm. Accessed September 26, 2016.

Figure 12. IDC illustration of percentages of useful data growth from 2013 to 2020.

The authors emphasize, however, that even of this useful grouping, "in 2013 perhaps 5% was especially valuable, or 'target rich.'" Characteristics of target rich data include information that:

- Is easy to access: data is easy to obtain and not locked away on individual PCs or in proprietary systems.
- Is available in real-time: Data is available when needed, and doesn't come too late to drive real-time decisions and actions.
- Has a footprint: The data has the potential to affect a large number of people and important parts of the organization.
- Is transformative: The data has the potential to actually change the organization or society in a meaningful way.
- Has intersectional synergy: The data has more than one of the above attributes.[71]

Is the metadata required to make the Smithsonian's digital files actionable over time complete and, at a minimum, available? And are those metadata treated as preservation objects that are maintained with the same level of care as the digital files with which they are associated? In many cases, it appears the answer is no. The Smithsonian's goal, then, should be to make all digital assets, that is, those resources that have enduring value, "target rich." Without doing this the vision laid out in the 2010-2015 Digitization Strategic Plan will not be realized.

In order to quantify the need, further research to fully understand the amount of unmanaged research data and other digital assets is required. A specific number is not necessary, but an accurate estimated range will help paint a realistic picture of the problem scope. Focus on research data first, but also look closely at collections that may be considered "rogue," as well. When evaluating collections, in particular, look holistically at the digital objects: what percentage are surrounded by a complete set of metadata and other information that will make those resources re-usable in the future? Finally, create projections that look at anticipated growth, not only current holdings, for both research data and collections.

[71] EMC and IDC. The digital universe of opportunities: Rich data and the increasing value of the Internet of Things — High value data. EMC2, April 2014. http://www.emc.com/leadership/digital-universe/2014iview/high-value-data.htm. Accessed September 26, 2016.

6.1.2 Communicate broadly

In order for the Smithsonian as an organization to get behind change initiatives aimed broadly at improving the long-term accessibility of digital resources, a campaign should be initiated to raise the level of awareness and galvanize staff to participate in the effort. This campaign should sound the alarm bell about the risks that Smithsonian digital resources face, and the potential impact of not acting.

The goal is to influence a change in behavior, and make all staff realize they have a role to play with regard to the long-term accessibility of digital assets. Most Smithsonian staff are already of the mindset that physical collections are everyone's concern, and caring for these is one of the most important functions of the Institution. Unlike physical collections, which are clearly held by a particular museum, archive, or library, responsibility for digital assets is unclear or they remain largely invisible to most staff, leading to the belief that digital stewardship is someone else's problem. The vision of the Digitization Strategic Plan is for everyone to take responsibility for the stewardship of the Institution's digital assets. Without the support of Smithsonian leadership, and a mechanism for communicating out the urgency for digital stewardship to the level of the individual, the long-term sustainability of digital assets will fall victim to the tragedy of the commons, and loss will be inevitable.

A communications plan — and acting upon it — should be a next step of the Digital Preservation Working Group, while additional elements of the program are being put in place. The urgency must be first recognized by Senior Leadership, then communicated throughout the organization from there with the Leadership's backing and support. Communicate the scale of the issue and the risks to the Smithsonian's digital resources if nothing is done. Position the problem of digital preservation as *everyone's problem.* Motivate people to want to participate and become change agents within their units, departments, or workgroups. Identify evangelists so that they can later be tapped to help with ongoing efforts.

6.2 Establish governance and oversight

Moving a change effort forward will require the participation of group of stakeholders who collectively have the authority to act, the expertise to advise, and the experience to lead. One of the most important contributors to the digital preservation challenges that exist today is the lack of an organizational mandate to ensure the long-term accessibility of digital resources, and lack of a structure to carry out such a mandate were it to be issued.

In order to fill current gaps in oversight, governance mechanisms must be established that are inclusive of all relevant roles, representing all digital resources of potential value. This includes the establishment a neutral entity that is responsible for coordination and oversight of all preservation efforts, advisory groups, and a definitions of roles and responsibilities throughout the organization.

6.2.1 Establish a Digital Preservation Directorate

The purpose and function of this office is to oversee the creation of a pan-Institutional preservation program, and guide its implementation over time. Operating similarly to the Smithsonian's National Collections Program, the Digital Preservation Directorate should be a neutral office that works collaboratively with all relevant stakeholders, including Smithsonian leadership, OCIO, DPO, units, and research centers and programs. This office should provide services including:

- Market and communicate the Institution's digital preservation strategy, tying preservation to other more established digitization and access efforts
- Recommend and enforce Institution-wide policy
- Advise individual units and groups on the development of local preservation policies
- Provide guidance and training to content creators, including researchers, curators, and collection managers on their responsibilities
- Liaise with service units, including SIL, SIA, unit IT staff, unit archives, and repositories to ensure that roles and responsibilities are understood and remain aligned
- Provide oversight and coordination of central technical services, ensuring that there is a managed environment to support all types of digital resources
- Plan for growth, and advocate for secure line-item funding to support long-term stewardship of digital assets
- Establish benchmarks for successful preservation
- Establish metrics and regular reporting data on preservation efforts to demonstrate progress and short-term wins
- Oversee periodic preservation audits of repositories and ensure follow through on necessary improvements

We recommend that this office have at least 4 staff: one director, one collections and institutional output coordinator, and two or more research data liaisons. Researchers will require additional support at the individual and lab level, and therefore a larger support staff will be needed to provide outreach and guidance to these groups and help them coordinate with other services across the Institution.

While the need for such an office has been recognized by digital preservation professionals in other organizations with similar missions, such as universities, our extensive research on this topic shows that these organization have yet to implement such an office with broad enough scope (i.e., only dealing with research data or library collections) or authority. The creation of the Digital Preservation Directorate will provide an important opportunity for the Smithsonian to demonstrate leadership in this area.

6.2.2 Establish an advisory board

The Digital Preservation Directorate should be supported by a diverse advisory group that is representative of preservation functions and different domains. The Digital Preservation

Working Group is an excellent starting point for this board, but will need to be expanded and reorganized to be most effective as initiatives are rolled out.

A simultaneous top-down and bottom-up approach to the establishment of this body will help ensure that transformation of business practices and processes are successful. Executive input is required to help prioritize goals, and provide validation and visibility to a shift in Institution-wide practice. The participation of existing digital resource stewards — those already working on this issue within their unit or service — as well as managers and researchers from across the organization will help ensure the interests of all stakeholders are represented. These two groups may decide to form a hierarchical council, in which executives sit on a steering committee, and other stakeholders form the official Digital Preservation Advisory Board. Working groups may be established for projects and tasks that require subject-matter expertise. Existing interest groups such as the Time-Based Media and Digital Art Working Group[72] can liaise with and report to the advisory board.

6.2.3 Define roles and responsibilities

The role of all central service groups, units, labs, and individual researchers must be defined in order to ensure that responsibility for preservation is shared. These roles and responsibilities should eventually be formalized in Smithsonian Directives, and should include the Digital Preservation Directorate, OCIO, Smithsonian Institution Archives, Smithsonian Institution Libraries, unit-level IT and collections managers, curators, lab managers, principal investigators, and research fellows.

Responsibilities should be detailed so that adequate support for all aspects of preservation are accounted for. This is particularly important for research centers, which may not have collection management staff they can turn to for help.

Each role should be made aware of the complementary roles throughout the organization and how their responsibilities dovetail. For instance, if SIL staff are tasked with providing support to researchers in the creation of data management plans and advising on data deposit, unit IT staff should be aware of this. Too often we heard from interviewees that it is unclear who to turn for help with different aspects of preservation, and depending on who they talked to, they would get different answers. By defining roles and then formalizing them in directives, processes can be significantly streamlined.

6.3 Create a vision for digital preservation

One of the first tasks of the advisory board will be to develop and communicate a simple, clear, and inspiring vision that illustrates the objective of digital preservation for the Institution, and couches it within the existing Smithsonian vision. This vision should be outcome-based and should capture the implicit preservation management component of reaching that outcome (e.g.,

[72] https://www.si.edu/tbma/about

that the digital resources the Smithsonian creates today are available to help our great-grandchildren make new discoveries). The vision for digital preservation should be tied to the vision for digitization — the end goals should be the same for both — but should be inclusive of the digital resources that do not originate from physical-to-digital conversion, namely research data and other content in born-digital form.

6.3.1 Demonstrate the value

The vision should tie to current outcomes that demonstrate the value of ensuring data is accessible and reproducible. For example, in the research domain, the authors of the Mathematical and Physical Science Open Data Report report note that, "There is ample evidence that the conscientious calibration, curation, and preservation of research data has immense benefits,"[73] and point to three significant examples, from the Hubble Space Telescope, Sloan Digital Sky Survey, and Laser Interferometry Gravitational Wave Observatory. All three generate and archive fully reproducible data, which have resulted in thousands of peer-reviewed publications by people who were not affiliated with the original investigations that produced the data, and some of the most significant scientific discoveries of the 21st century.

Similar stories can be found across the Smithsonian to illustrate the value of both collections and research preservation. Collect and tell those stories. Where they do not exist today, talk about the potential, and how to make them a reality. These outcomes are why the White House Office of Science and Technology Policy[74] issued its policy on public access to federally-funded data, and why the NSF[75] and other agencies have made data management a fundamental component of grants.

6.3.2 Incorporate the vision into Strategic Plans

It is important that all stakeholders across the Institution support the vision and see it as a shared goal. Codifying it into the Institution's Strategic Plan is a critical first step. The Smithsonian's current strategic plans will reach the end of their term in 2017, which means now is the time to start incorporating the vision, goals, and objectives for digital preservation into the next plan(s). While both the 2010-2015 Strategic Plan and the Digitization Strategic Plan laid important groundwork for the creation of a Digital Smithsonian, they did not go far enough to anticipate what would be required for full life-cycle management of digital resources. They also

[73] Hanish, Robert, et. al. MPS Open Data workshop series draft report. MPS Open Data. https://mpsopendata.crc.nd.edu/images/Reports/MPS_ReportDraft_v4.pdf. Accessed September 26, 2016.
[74] Holdren, John P. Increasing access to the results of federally funded scientific research. Executive Office of the President, Office of Science and Technology Policy, February 22, 2013. https://www.whitehouse.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf. Accessed September 26, 2016.
[75] National Science Foundation. NSF 15-1: Chapter II - Proposal Preparation Instructions, December 26, 2014. http://www.nsf.gov/pubs/policydocs/pappguide/nsf15001/gpg_2.jsp. Accessed September 26, 2016.

fall short in ensuring that digital research data can play the role it is envisioned to — contributing to a body of knowledge and facilitating new discoveries — over the long-term.

If the next planning cycle produces a digital plan (similar to the Digitization Strategic Plan) — and we recommend that it does — it should focus equally on digitization, born-digital resources, description and tagging, and long-term stewardship. The goals and objectives of this plan should clearly highlight what is required to reach system-wide preservation of the valuable digital resources held by the Smithsonian.

6.4 Create and update policies for digital preservation

The lack of clear digital preservation policies contributes to confusion and sometimes paralysis when staff are faced with determining how to appropriately manage of the Institution's digital resources.

6.4.1 Formalize terminology

Lack of formal terminology surrounding digital resources, their accessibility, and their management results in a great deal of confusion, and sometimes conflicting interpretation of existing policy. Although it can be argued that content is generally more important than the form it takes (i.e., the text of a book has more inherent value than the book as a physical object), digital content requires specific rules for managing its long-term accessibility, and all parties who support these resources must share a common understanding of them. Clarity is required to provide consistent direction on the care of digital collections, digital associated information (e.g. metadata), digital research data, and digital Institutional output. These terms may be commonly used to describe the digital content held by the Institution, but they may have different meaning for different people or institutional units. Definitions should be clear, and consistently applied; relevant policies should be updated to reflect them.

For collections, SD 600 should be updated to include language that helps define digital collection items (PDCO, digital surrogates) and metadata. Define what "associated information" means in the digital age and when this information should be considered part of, and stewarded as, collection resources. Outline when collections conservation data should fall under SD600. SD 610 defines several relevant terms that can be used as a starting point.

For research data, define the categories of data, using those provided in this report as a starting point, and the types of data that comprise them (e.g., raw data sets, logs, white papers, codebases, or publications). Define who a "researcher" is — curator, principal investigator, staff scientist, post-doctoral fellow, or all of the above? Provide guidance on when research data may become collection items. For example, can a 3D representation of artifacts in the field be accessioned as a voucher in lieu of the physical objects? Define data reproducibility and the requirements to enable this at the discipline level.

For institutional output, clarify what types should be stewarded as digital assets. Provide clear guidelines that helps delineate when a digital resource created by Institution staff should be accessioned as a type of collection item, and when it should not.

6.4.2 Create a Smithsonian Directive for Digital Preservation

As illustrated throughout this report, SD 610 does not go far enough to define a complete set of preservation management roles and responsibilities. Because the focus of this directive is on digitization, the preservation components that are addressed within it exclude actions required for the long-term management of research data and PDCO. This leaves a gap in digital preservation policy across the Institution, which must be corrected in order for systematic, Institution-wide preservation to be adopted as practice.

We recommend that a new directive be issued that outlines institutional responsibilities and individual requirements for digital preservation at all levels of the Institution. It should follow the outcome of the strategic plan (see Section 6.3.2 above) and dovetail with SD 610 for digitization and SD 609 for access. The directive should include — or reference from other directives — definitions for terminology that is not clear (e.g., digital assets, associated information, digital preservation) or that may have domain-specific meanings (e.g., research data, digital collections). It should formalize the decisions made during the governance creation process, when preservation roles and responsibilities of various groups from across the Institution are defined.

6.4.3 Create a Smithsonian Directive for Research Data Management

Researchers are motivated and incentivized by the academic need to publish, of citation, and of funding awards to further their work. Not surprisingly, the ways that the data they produce is managed is subject to those same incentives. As such, it may be deposited in third-party, discipline-specific repositories, or with collaborating Institutions such as universities, where they might have more visibility within their field. In fact, in some cases, they are being encouraged by their units or fellow researchers to deposit data in these external repositories, both because it may be in their best interest professionally, and because the Smithsonian doesn't always have the infrastructure or staff to preserve data on their behalf.

This may actually be acceptable to the Smithsonian, however, whether or not it is cannot be determined because there are no formal policies, and no formal Institutional commitment made regarding the preservation of research data. Before moving forward, the Institution must decide what its position on research data management is, answering questions such as:

- Does the Smithsonian commit to preserving Primary / Raw data, but not Analyzed / Derived data?
- Does it only support fully packaged, documented, and reproducible research?
- If a researcher wants to deposit data in an external repository, does the Smithsonian want a record of that? What should that record look like?

- Does the Institution provide guidelines on selecting repositories?

These questions and others must be clarified at the executive level before policies can be issued.

Once the organization's commitment is determined, it should be formalized in a Smithsonian Directive that details scope, definitions, roles and responsibilities, and Institutional mandates for the management of research output. The policy should provide support toward compliance with federally mandated access and management requirements, and researcher's incentives for publication, while also aligning with the Smithsonian's own interests toward the preservation of research output.

In the near future, it is likely that funding agencies will ask for details about the support to be provided by the researcher's institution in grant applications.[76] To bridge researcher's goal of identifying a means for managing research data, and the institution's mission to ensure re-usability of this data, JISC has recently published a guide for "Developing an organisational profile for research data management services," which provides a checklist of areas of support organizations that employ research staff should visibly account for, including: research data management policy, advice and support services, storage, data registry or catalog, data access procedures, and more.[77] Organizations can use this list to identify internal areas of improvement, and once complete, availability of this information in a single location can be enormously helpful to researchers, who, as noted, don't always know where to turn for help. The Smithsonian may want to consider using this guide when developing its own research data services.

6.5 Establish mechanisms for enacting organization alignment and accountability

Smithsonian Directives are imperative for providing an Institution-wide context, and for outlining the mandate, purpose, and policies, for functions that impact the entire Institution. Issuance of a digital preservation directive will be an important first step to ensure that digital resources will be effectively preserved, but the work does not end there. Interviewees repeatedly stressed the importance of accountability, feedback, and oversight for those responsible for interpreting and implementing local policies based on directives. In order to be effective, there must be an effort to directly align and engage the individual staff and units of the Institution to a high-level preservation directive.

[76] In the UK, the Engineering and Physical Sciences Research Council is already requesting this as part of grant applications.
[77] Davidson, Joy. 'Developing an organisational profile for research data management services - a guide for HEIs'. Edinburgh: Digital Curation Centre, 2015, 9. http://www.dcc.ac.uk/projects/opd-for-rdm

6.5.1 Create a pan-Institutional digital preservation vocabulary

Once formal terminology is established and codified in a directive, individuals and units must be able to interpret and map it to their own context and domain-specific vocabulary. The Digital Preservation Directorate and members of the advisory board will need to provide outreach and assistance at all levels of the Institution, to relate the Institution's digital preservation vision, strategic goals, and policies, to local practice, and to empower all stakeholders with the means to act in alignment with them. To do this, the Directorate will need a deep understanding of the various disciplines of research and units engaged in digital collections creation, how information is being communicated to them from their communities, and the changes being initiated by funders, publishers, and peers, so that their unique vocabularies and practice can be mapped to the Institution-level policies and guidelines. As external forces change (i.e. federal funding guidelines), or systems within the Institution improve and grow (i.e. robust research data repositories), the message being transmitted between the Directorate and local practitioners must be adapted. The Digital Preservation Directorate should maintain this documentation, make it readily available (e.g., on a website) and update it as change occurs.

6.5.2 Conduct training and outreach

A key component of this digital preservation strategy is a programmatic and coordinated organization-wide outreach, education, and capacity-building initiative. Unit directors, collection management staff, lab managers, and principal investigators will require training in order to enact local policies and ensure consistent compliance with new mandates.

A formalized program dedicated to providing content creators with guidance about digital preservation concepts, individual responsibilities, data management plans, and other topics, is necessary to move policy into practice. A baseline measure of success will be for staff to know when they don't know something and where they can go for help. Outreach and training will need to be a combined responsibility of the Directorate and unit-level staff who are trusted by their constituents and are best able to understand the specific context, needs, and goals of the content creators. Official training guidelines should be created and maintained by the Digital Preservation Directorate, and communicated to experts and leaders at the unit or program level, who understand the Institution's vision and strategy for digital preservation, and can provide discipline-specific guidance.

Outreach would include workshops, tutorials, and consulting for existing staff, as well as new staff and research fellows as part of their onboarding process.

6.5.3 Create accountability structure for enactment of policy

Developing and issuing policy is step one, but auditing adoption must follow closely behind. Verifying that interpretation and application of policies across domains is consistent and in accordance with Institutional goals helps to engage local units with the broader operation and identify when a unit needs support. Ongoing communication between content creators and

central services supports oversight and validation of local plans, and administering deterrents when practices place digital assets at risk.

Of course, any approach that uses a stick must be accompanied by a carrot. Incentives for compliance, including infrastructure, staff support, and funding, are described in recommendations below.

6.5.4 Establish, track, and report on metrics that illustrate the value of digital preservation

Tracking digitization progress on the Smithsonian Institution Dashboard is a simple way to demonstrate progress, but it also points to the scale of the work that remains to be done. Expanding this page to include metrics that illustrate digital preservation goals could provide meaningful data and enforce Institutional accountability to both internal stakeholders and the public at large. Metrics might highlight outcomes related to the accessibility, understandability, and repurposing of data, such as:

- Total amount of data in digital preservation environments
- Total research data available to the public
- Amount of data repurposed for new research
- Number of citations of publicly accessible datasets
- Number of peer-reviewed publications based on data generated by Smithsonian researchers

6.6 Ensure supporting technical infrastructure

New requirements and responsibilities for digital preservation must be matched by a robust technical infrastructure that can support the digital assets that the Institution commits to govern. Without this, units, labs, and principal investigators lack an important mechanism for compliance, and their incentives to contribute this data to the scientific arene in the future succumb to more pressing day-to-day responsibilities. By offering an infrastructure that meets the needs of these groups for management after the completion of projects, at minimum, the Institution is better able to gather and track data, and the content creators are provided an important resource that helps them perform, maintain, and pass on their work.

In some cases, this might require expanding the role of existing repository services, but in others it will necessitate building out new, actively managed storage environments. It also might mean that the Institution recommends that external services be used, such as domain-specific repositories, but that it tracks where datasets reside so that interventions can be made if, for instance, an external repository ceases to operate.

The Digital Preservation Directorate may also explore participation in emerging collective preservation networks, such as the Digital Preservation Network (DPN).[78] While such services are still emerging and not currently a viable solution for the breadth of the Smithsonian's digital assets, the Institution's participation in these efforts today may help drive them forward. Further research is required to determine the suitability of such a service to the Smithsonian.

6.6.1 Clarify the role of existing repositories

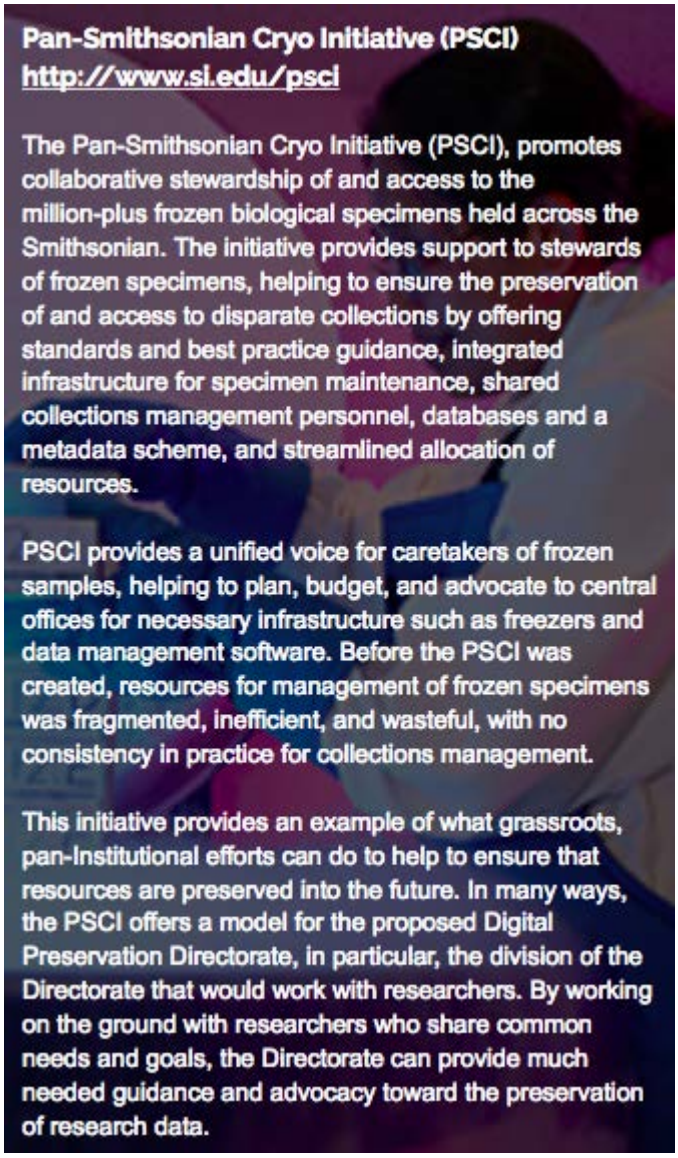
During our research, we examined four repositories that offer services to support content creation, access, and/or preservation. When these very different technological responses to the challenge of digital asset management are viewed together, it becomes clear that there is no central, supporting technology or administrative infrastructure for the various data types the Institution that may have preservation potential. The Institution must make clear its commitment to preserving assets by investing in a robust infrastructure that is available to all content creators as their needs mature.

We recommend that each existing repository establish its scope. Where collecting gaps are exposed that leave resources without an obvious "home," the repository managers, in coordination with the Directorate, must decide whether to update existing policies to include these resource types, or recommend the creation of new repositories that will take responsibility for them.

For example, most digital collections assets are held in the DAMS, as well as a significant amount of other institutional output, such as event video and photography, marketing collateral, and exhibition resources. However, the DAMS scope is currently defined by resource type — image, audio, video, or time-based media art — which leaves out a number of digital collections types. If DAMS is determined to be the technical solution for collections, there must either be a mechanism for expanding the DAMS to support new data types such as 3D items, email archives, and Microsoft Office documents, or a separate repository dedicated to specialty formats and large collections (the role that SIA Digital Archives plays now for its own collections), should be formalized. It is essential that clarifying and filling these repository collecting gaps be made a priority.

While we don't necessarily advocate that repositories be expanded beyond their realistic capabilities and stretched too thin to be effective, we feel strongly that technologies should support the requirements of the different stakeholders and evolving content types. It is inevitable that new data types of increasing complexity will be collected and created. The Smithsonian needs a technical infrastructure that can respond to these changes.

[78] http://dpn.org/



Infrastructure isn't limited to just repositories, however. Short-term storage is also needed during the creation process (e.g., research or digitization). Often this type of storage requires a low barrier to entry and worldwide access, which is why many stakeholders favor easy-to-use, familiar, third-party tools like Dropbox, or external hard drives that can be passed between researchers. Regardless of the technology or services used for short-term storage, fluid pathways must be established that will move important digital resources off of those storage media to permanent, managed storage. Motivating stakeholders to do this will likely require both carrots and sticks.

6.6.2 Gather requirements for research data infrastructure

Once the organization determines its preservation commitment to research data, financial and administrative support to build out a technical infrastructure must closely follow to stem the tide of short-term, project-based solutions that rely on time-limited funding that does not consider the longevity needed to sustain digital content. In order to determine what technological support content creators need to manage their data over whatever time period, it is essential to perform high-level requirements gathering by seeking input from a large sample of stakeholders. This will ensure that solutions are not created in a vacuum, and consider the variety of content being produced across the Institution. Care should be taken to determine where there should be dedicated infrastructure (e.g., for genomics, space science data), and where they can be shared (e.g., biological sciences).The Pan-Smithsonian Cryo Initiative provides an example of a grassroots model for gathering and communicating the collective requirements of researchers who work with frozen biological specimens, which may offer insight to this process.

We recognize that SIdora is a current OCIO effort to provide an infrastructure for active research. However, the input gathered during our interviews indicates that this repository may not be meeting current needs. We recommend that rather than continuing to build out this service, pause development and initiate a comprehensive and coordinated requirements

development process. The outcome of this effort may indicate that SIdora needs to be re-imagined, re-branded, or re-architected. There is the potential, as well, the research finds that a new system should replace it.

6.7 Operationalize digital preservation funding

In addition to defined roles and responsibilities and technological infrastructure, committed financial support is required to maintain digital resources over time.

6.7.1 Establish line items for preservation support

To cement the value of digital preservation at the Smithsonian, programmatic funding for staffing, technology, and related resources should be represented through requisite budget lines. Budgeting for a growing cadre of staff in various parts of the Institution to carry out new preservation roles and responsibilities will be necessary — reliance on project-based funding may help to build systems initially, but will leave the Institution without expertise to maintain them over time. Central services, for example, will require new staff roles to support expanded mandates and provide ongoing guidance, training, and oversight to staff across the Institution. Units and programs will need local resources to help ensure participation in the preservation program, and technological infrastructure implementation, growth, and maintenance must be considered at the fore.

Once new infrastructure is tested, piloted, and rolled out, ample IT and Digital Preservation Directorate support must be provided to facilitate use. Adoption of centrally funded and supported technologies will be key to fulfilling an organizational goal of ensuring preservation, but it can't be expected that the users will do everything needed to make digital assets usable over time. As DAMS staff can attest, it is critical that some staffing resources are dedicated to support content creators, and to ensure that the digital resources and associated metadata being created are compliant with system requirements for access and preservation.

6.7.2 Move away from reliance on project-based funding

Sustainability and persistence requires that digital preservation programs have ongoing financial support; short-term funding has been proven to leave digital resources vulnerable and subject to loss. Reliance on project funding for research data management, for example, has already contributed to a significant backlog of unmanaged, and therefore unusable (and often un-findable), research data. The size of the grant does not matter — rarely do external funding sources support sustainability of research data beyond the lifetime of any grant. When funding runs out, researchers do not or cannot (often due to cost or lack of expertise) keep their data storage up to date and accessible.

A sustainable approach should combine programmatic and project-based funding. A small percentage of project funding should be allocated toward data management, becoming part of the overhead the Institution likely requires of grant applications already. These funds should be

contributed to project-specific technological and staff support, which is matched through ongoing programmatic funds. Funding for the program should ensure that technologies are maintained and updated over time, that policies and procedures are documented and updated, and for ongoing efforts such as outreach and training. Together, these funds support the preservation, publication, and long-term access to the research. This approach is more likely to be successful providing that the Smithsonian describe what benefits it provides to the longevity of the project, so that researchers understand what they are getting for their money.

6.8 Implement a phased approach

> *Sustainable preservation strategies are not built all at once, nor are they static. Sustainable preservation is a series of timely actions taken to anticipate the dynamic nature of digital information.*
>
> *— Blue Ribbon Task Force on Sustainable Digital Preservation and Access*[79]

Finally, we need to recognize the scale of this undertaking, and that even though the quantity of digital content continues to grow on a daily basis, it is impossible to take on all of the challenges — past, present, future — at once. Planning for the future is the first step — building systems and methodologies that are forward-looking prepare the Institution for the massive growth that is estimated over the next ten years. Once future-proof systems are stable and can begin to capture content created in the present, the Institution can programmatically approach solutions for the enormous data backlogs that are at risk of loss. This approach distributes resources in a sustainable way, and provides time to develop systems and storage that can accommodate the backlogs. As success is demonstrated throughout each phase, and aspects of the program are tweaked and improved, the backlog can slowly be chipped away.

Changes do not occur overnight, but take a commitment of and investment in technical, financial, and human resources, and support from Institutional leadership to ensure these investments happen. With time and directed focus, coordinated and sustained systems will begin to yield desired outcomes that will allow the Smithsonian's vision for a digital tomorrow to be realized. With sustained care of digital resources, collecting, digitization, and research will bear fruit by providing content to support exhibitions, education, access, publication, and more, across the Institution. This investment in our digital heritage and the scientific record is not being made once, but is renewed with each new initiative, new researcher, and the desire of every user to access the Smithsonian's collections no matter the format. Laying the foundation now will position the Smithsonian to respond to growth and change with the flexibility required of a digital leader.

[79] Blue Ribbon Task Force on Sustainable Digital Preservation and Access. Final report of the Blue Ribbon Task Force on sustainable digital preservation and access, February 2010. p. 5. http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf. Accessed September 28, 2016.