STEWARDING THE INVISIBLE

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

Executive Summary
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.

EXECUTIVE SUMMARY

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 institutional resources. Realizing this vision would require digitization of collections, integration of digital research output, and overall improvements to the description and accessibility of all digital resources.

In the ensuing years since the Strategic Plan was published, a great deal has been accomplished toward these goals. However it has been observed that the digital preservation component of the strategic vision has not been addressed formally, and that today, newly digitized and born-digital resources are at great risk. 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.

In 2015, former Secretary of the Smithsonian, Wayne Clough, charged a pan-Institutional Digital Preservation Working Group (DPWG) 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 preservation 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. Data for this study was gathered through interviews with 16 stakeholder groups from across the Institution, an online survey of 100 Institutional researchers, and review of existing policies, strategy, and other documentation.

The report seeks to identify *what* gaps exist in current digital preservation practice, *why* these are occurring, and *how* the Smithsonian can make improvements to ensure the preservation of its digital resources.

Findings

After careful analysis of data gathered during this project, we have reached several conclusions that can be summarized as follows:

- Existing policies that address digital resources focus on digitization, and are ambiguous with regard to roles, responsibilities, and tasks for digital preservation.

- Existing policies (e.g., SD 610) put preservation responsibility solely in the hands of the resource creator, and do not specify any central roles to support these stakeholders.

- The Smithsonian does not offer formal, central preservation support to creators or stewards in the form of guidance, oversight, infrastructure, guidelines, policies, or staff. The resources that do exist, such as the DAMS, play an *ad hoc* preservation role.

- Content creators (e.g., researchers) and collecting units clearly state that taking on full preservation responsibility in addition to primary responsibilities is an unrealistic expectation. Furthermore, most do not have the expertise to ensure that digital resources will remain accessible over the long-term.

- Because there is no mandate for digital preservation support (and no organizational body to offer it), creators of digital resources are unsure where to turn for help. For many, storing data on at-risk media is the *de facto* response. As a result, there are large volumes of unmanaged data today.

- The current project-centric approach to data management is inefficient, wasteful, and places digital assets at risk of loss because administrative resources are temporary and funding is limited.

- The volume of potential digital assets that require long-term management is enormous. There is an estimated 6.7 PB of research data scattered across the Institution, and nearly 2 PB of collections data and other institutional output. These numbers are anticipated to increase dramatically in the coming years.

- There is a lack of a mechanism to quantify the scale of existing digital resources, and the pace of growth has been paralyzing.

- The lack of clarity around definitions of digital resource types (i.e., what is the scope of “digital collections” and “research data”) inhibits policy creation and designation of responsibility.

- There are functional and storage capacity gaps in the existing infrastructure if all digital resources of value are to be managed by the Institution over the long-term.

- Without designating responsibilities for all aspects of digital preservation, complacency will persist.

The current situation results in silos of effort across the Institution. Excellent groundwork has been laid by the dedicated staff who manage several repositories and asset management systems, and collecting units who have taken great care to ensure their resources are properly managed. 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—they are unidentified, unmanaged, and as time passes, they face the very real prospect of loss.

Recommendations

Just as it is investing in digitization, the Smithsonian must invest in the creation of a digital preservation program, which will ensure the sustainable preservation of digital resources through coordinated effort. Tasks toward this end include the following:

1. Instill a sense of urgency for the need for digital preservation amongst stakeholders. Quantify the need and communicate it broadly.

2. Establish governance and oversight through the creation of a Digital Preservation Directorate, a Digital Preservation Advisory Board, and definition of cross-institutional roles and responsibilities from the Secretary down to the level of the individual researcher.

3. Create a vision for digital preservation that demonstrates the value and impact that preservation can have. Incorporate that vision into the next Strategic Plan.

4. Create and update policies for digital preservation. Start by formalizing and updating outdated terminology, then create a Smithsonian Directive that specifically addresses digital preservation, and a Smithsonian Directive specific to research data management.

5. Establish mechanisms for ensuring organizational alignment and accountability, such as through creation of a pan-Institutional vocabulary that maps formal terms to specific contexts, training and outreach, policy accountability framework, and by tracking metrics.

6. Ensure that requisite technical infrastructure exists to support preservation needs for researchers and collecting units. Clarify and/or expand the role of existing repositories. Create a mechanism for researchers to collectively communicate technical requirements.

7. Ensure sustainability of the digital preservation program by formalizing ongoing funding streams that are supplemented with project and grant funds.

8. Implement a phased approach to rollout the new program so that data is appropriately captured moving forward, while methods for dealing with existing backlogs are established and executed.