**Research Data Management Program Pilot**

REPORT AND RECOMMENDATIONS – EXECUTIVE SUMMARY

2018-05-04

## Background

The Research Data Management Program Pilot (RDMPP) was a six-month focused effort to devise and model a programmatic, sustainable approach to managing research data at the Smithsonian. The pilot included staff from the Smithsonian Libraries (SIL), Smithsonian Archives (SIA), and the Office of the Chief Information Officer (OCIO). The primary deliverable of the pilot is this report, which includes detailed recommendations on what technical infrastructure to adopt; what services to offer; how to structure a RDM program; and draft policies and best practices. The recommendations in this report are distilled from previous SI studies related to research data, interviews with SI researchers, a review of comparable policies and programs at Federal agencies and Universities, and a review of current literature around RDM. Specifics for implementation are provided in the appendices.

Over the past eight years, the Smithsonian Institution (SI) has explored limited RDM servicesi at both the enterprise and unit level.ii None of these efforts, however, have led to a sustainable program with sufficient resources to support management of the digital research output of the Institution. A recent studyiii estimated that there may be 6.7 PB – the equivalent of nearly 1.5 million DVDs worth – of research data at the Institution, with less than 13% stored in a safe, actively managed environment, and an unknown percentage publicly available for reuse.

Like the National Collections, the Smithsonian’s research data are assets of enormous value that reflect the breadth and depth of scholarship at the Institution. Often created with public funds, the Smithsonian has an obligation to steward these assets responsibly and share them broadly. However, in the absence of clear institutional mandates and dedicated resources, the burden of managing research data has largely fallen on individual researchers – a “small science” iv approach that is inefficient and untenable. The Institution now has an opportunity to embody the ideal of One Smithsonian to collectively address our responsibilities towards these research assets, and as a result increase our reach and impact far into the future. Without an effective research data management (RDM) program, however, research assets are at increasing risk for future loss – a loss to the Institution, the scholarly community, and the world.

“[it is] increasingly anachronistic to continue the small science approach to data management, because it inhibits the re-use and integration of data….Data re-use has become particularly relevant in the face of global environmental challenges whose solutions call for deeper understanding of complex systems across multiple disciplines, places, and time periods.” v

## Recommendations

To prevent data loss, increase the visibility, reach, and impact of the Smithsonian’s scholarly work, and enable our researchers to more easily comply with existing Federal and Smithsonian directives, the Institution should establish a centrally administered research data management program. The program should be backed by sound policies, effective technical infrastructure, and sufficient human and financial resources to support the management of research assets at all stages of the data life-cycle. The program should leverage and build upon existing efforts in the units to develop a pan-Institutional Research Data Management network that works cooperatively to achieve the goals of the Institution.

1. [Define and institute policies that support FAIRvi data principles](#_Define_and_institute).
2. [Establish a central research data program office.](#_Establish_a_central)
3. [Provide cooperative services that support FAIR data practices.](#_Provide_cooperative_services)
4. [Strengthen technical infrastructure](#_Strengthen_technical_infrastructure).

As immediate **next steps** towards creation of a research data management program, the Institution should:

1. Approve the charter for and **form a** **policy working group** that will define “research data” for SI and detail roles and responsibilities around their management.
2. **Establish a central** **Research Data Management Program Office**, modeled on the Smithsonian’s National Collections Program.
3. **Identify resources** to staff the Program Office, including data experts to embed in units, as well as seed a RDM “pool” type fund.
4. **Quantify at-risk research data**, and facilitate infrastructure planning, by conducting a formal inventory.

Until a central program office can be established, and while the policy working group is being assembled, SI should **maintain momentum** towards an RDM program. OCIO, SIA, and SIL will need to devote resources to moving the Institution forward by providing interim services and initiating projects such as:

* Complete repository testing against requirements and set timeline for launch.
* Include RDM infrastructure needs into any future technology planning, e.g., pipelines for very large file transfer, storage.
* Plan early CY19 RDM share fair, identifying any funds needed.
* Expand training opportunities, including cross-training, for staff in RDM best practices.
* Have pilot team members participate in the Digital Library Federation eResearch Networkvii (May-October, 2018)
* Begin collecting dataset citations in SRO to provide metrics to demonstrate research impact.

Read the full report: <https://si-confluence.si.edu/display/RDMPP/>

1. RDM services can be defined as “providing information, consulting, training or active involvement in: data management planning, guidance during research….documentation and metadata, sharing and curation (selection, preservation, archiving, citation) of published data” –ARL SPEC kit 334.

Efforts include development of SIdora platform and the creation of the Office of Research Information Services (now the Office of Research Computing) in 2011, the 2011 OPandA study on sharing research data in biology, and SIL’s Data Management Working Group, formed in 2013.

AVPreserve. 2016 (See [Appendix E](#_Appendix_E:_Excerpts))

Here “small science” refers to data management practices driven by the needs of individual, small-scale, research projects.

OPandA. 2011

[FAIR](#_Glossary_of_abbreviations) is Findable, Accessible, Interoperable, and Reusable <https://www.diglib.org/opportunities/e-research-network/>