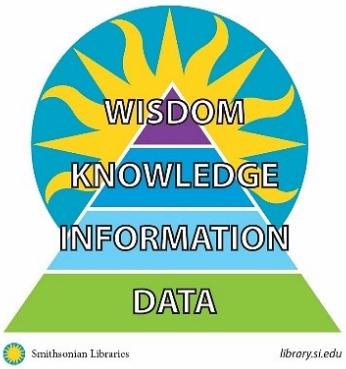
Research Data Management Program Pilot

REPORT AND RECOMMENDATIONS

Prepared by
Keri Thompson, SIL
and the Research Data Management Program Pilot Team
Lynda Schmitz Fuhrig, SIA
Beth Stern, OCIO
Sue Zwicker, SIL

2018-05-04



EXECUTIVE SUMMARY

Background

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"[i] 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.

Over the past eight years, the Smithsonian Institution (SI) has explored limited RDM services[ii] at both the enterprise and unit level.[iii] 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 study[iv] 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.

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 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.

> "[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

In order 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 FAIR[vi] data principles.
2. Establish a central research data program office.
3. Provide cooperative services that support FAIR data practices.
4. 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 Network[vii] (May-October, 2018)
- Begin collecting dataset citations in SRO to provide metrics to demonstrate research impact.

---

i. OPandA. 2011. (see Select Bibliography and Appendix E). Here "small science" refers to data management practices driven by the needs of individual, small-scale, research projects.
ii. 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.
iii. 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.
iv. AVPreserve. 2016 (See Appendix E)
v. OPandA. 2011
vi. FAIR is Findable, Accessible, Interoperable, and Reusable
vii. https://www.diglib.org/opportunities/e-research-network/

A Series of Unfortunate Events

A research group at SI was conducting destructive testing using samples from irreplaceable specimens. When their manuscript had been accepted for publication they were required to submit their raw data with the publication. By that time the lead author, a Post-doc, had left the Smithsonian. Though the lab had procedures on where and how to save data and descriptions, when a fellow researcher went to retrieve them from local storage they were not there. The instrument used to conduct the experiments, which would have had a copy, was no longer on site. The only clue to find the raw data (if indeed they were anywhere) was the filenames of the analyzed data, a cryptic combination of the first author's initials and an arbitrary sample number that didn't correspond to the sample names used in the experiment.
Ultimately, the raw data file names were painstakingly deduced using one piece of metadata found in each analyzed data file. Those filenames were then used to locate the original raw data stored amidst hundreds of other files on a second backup disk the first author eventually found. In all, it took three researchers nearly two weeks to recover the data that had taken several months to create.

What if the second backup hadn't been found? What if the analyzed data files had no metadata automatically inherited from the raw data?

Alternately…what if the data had been properly backed up in three places? What if the first author had followed good file naming practices?

TABLE OF CONTENTS

INTRODUCTION

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 the American people to steward these assets responsibly and share them broadly. Managing digital research data in a way that enables future access and re-use serves not only to fulfil the mission for the "Increase and Diffusion of Knowledge" but can strengthen the public trust in the scholarly process and its outcomes by promoting reproducible research and verifiable facts.

Background

Over the past eight years, the Smithsonian Institution (SI) has explored limited research data management (RDM) services at both the enterprise and unit level. 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. The Research Data Management Program Pilot was proposed in 2017 as a short-term focused effort to model a sustainable, programmatic approach to working with research data. The pilot would serve as a catalyst to begin moving the Institution away from the current ad hoc management efforts surrounding the Institution's research output, and suggest concrete next steps for developing a programmatic approach compatible with the strategic goals of the Institution. With staff from three pan-Institutional units (OCIO, SIA, and SIL) the pilot was tasked with producing draft policies and best practices as well as recommendations for program administration and structure, program services, and technical infrastructure.

The existing resources that support RDM at SI are both underutilized and insufficient to support a majority of the research output of the Smithsonian throughout the full data lifecycle. Generally, researchers have limited time and capacity to devote to each phase, so tend to focus on those directly connected to production of peer-reviewed publications, including collecting, analyzing, and sharing data, and to a lesser extent planning and documenting. Long term curation[1] and preservation, of necessity, must be carried out by an organization rather than individuals.

> Our results reinforce the notion that, in the long term, research data cannot be reliably preserved by individual researchers, and further demonstrate the urgent need for policies mandating data sharing via public archives.[2]

In order to make SI data findable and reusable long into the future, the burden of managing that data must be shifted from individual researchers to a team that includes researchers and data and information professionals. By working cooperatively RDM activities become more efficient through economies of scale and implementation of standard workflows and practices. A team approach also increases the potential value of the data itself – independent review can ensure data is described well enough to verify and reproduce results, or enable reuse in novel ways to create more knowledge.

[1] Digital curation involves maintaining, preserving and adding value to digital research data throughout its lifecycle – this includes typical archival activities such as appraisal (determining which pieces of data should be kept for the long term), disposition (removing objects not selected for long term curation), and transformation (migrating data from one format to another).
[2] Vines et al. 2014

"The lack of control by the Smithsonian … for the vast majority of data being generated by researchers is a significant potential risk for the loss of data." [3]

There are real risks in not addressing the stewardship needs of research assets. A recent study[4] 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.

In the most serious situations there are legal implications for loss of research data, such as those associated with experiments related to evidence in civil or criminal legal cases. For some classes of observational data, e.g., weather measurements, the raw data cannot be recreated, and if lost are lost forever. There is also a risk to the reputation of the Institution if data is not available (or is too poorly documented) to confirm research findings – the so called "replication crisis."[5]

Every delay in programmatically addressing RDM at the Institution only increases the need for future resources to tackle the growing backlog of data. The speed at which research data is being generated, media is degrading, and older formats are becoming obsolete is only accelerating. Metagenomics sequencing generating files of multiple gigabytes, and the need to support high resolution images from drones and streaming sensor data increasing both the volume and pace of data ingest. Raw data produced on equipment from the 1990s may not even be read-able, and thus re-analyzed, today.

The intangible losses caused by inaction, however, can never fully be known. These are the missed opportunities for future re-use of research data that limits the full potential of the Institution's scholarly outputs.

Previous Studies

The Research Data Management Program Pilot drew heavily from two previous internal studies related to sharing research data and developing digital preservation capacity, and has re-affirmed the findings of those earlier reports. A summary of the conclusions from those reports appear in Appendix E.

In 2010-11, the Office of the Under Secretary for Science commissioned a study from the Office of Policy and Analysis (OPandA - now called Smithsonian Organization and Audience Research) on how the Smithsonian shares digital scientific data in biology. Though the study was exhaustive, and made only three core recommendations, few of the

Legacy Data

A researcher at the Smithsonian Environmental Research Center was preparing to retire, and wanted to make their observational data publicly available for others to reuse. The data had been collected and analyzed by a post-doc who had since left the institution. The data was stored in thousands of spreadsheets without a "codebook" or dictionary that described the column headers, which included labels such as "S" (and nothing else.) All the files were given to the Office of Research Computing for loading into the SIdora data repository. For anyone outside the researcher's specific field of study, the contents of the files were effectively meaningless without a codebook, but because the post-doc had moved on to another institution, the descriptions of the contents of the files was not a priority.

Eventually, SERC was able to hire an ecology post-doc to describe the files' contents so they could be deposited and shared.

If SERC hadn't been able to hire a subject matter expert that understood the data, what future insights might have been lost because someone in another field, hoping to do cross-disciplinary work, couldn't interpret and reuse those data?

[3] AVPreserve. 2016 p. 52
[4] AVPreserve. 2016
[5] Baker, Monya. 2016. *1,500 scientists lift the lid on reproducibility*. Nature News. doi:10.1038/533452a

recommendations and next steps received follow-through. The majority of this study's conclusions and recommendations remain relevant, and have been liberally incorporated into this document.

In 2015-16 the Smithsonian Digital Preservation Working Group, established by Secretary Clough, commissioned a digital preservation readiness assessment. During the assessment, it became clear that there is a critical need to support storage and sharing of research assets during the active phase of their lifecycle to enable long-term preservation. Several conclusions and recommendations from the final report written by consultants from AVPreserve, *Stewarding the Invisible,* have been incorporated into this document.

Methods

The pilot approached developing a programmatic model in several ways:

- Researching RDM program structures at comparable Universities and Federal Agencies.
- Investigating best practices and standards through literature reviews.
- Reviewing previously published internal studies related to data management needs at SI.
- Gathering feedback from researchers through one-on-one and small group interviews.
- Trialing education initiatives.
- Doing market research into data repository systems, creating functional requirements, and testing systems.

Scope

The pilot chose to focus on the most immediate RDM needs at the Institution – currently active research and data, primarily in the sciences. The assumption is that solutions for pressing current issues would be applicable to, if not all, at least a majority of emerging issues in RDM such as Digital Humanities data, or legacy issues, such as research stored on outdated media. Those related topics were not completely ignored, however, as researchers do see a need for responsible management practices – particularly with legacy data – and some did express concerns to that effect. Once a programmatic approach to RDM is adopted, a more complete picture of the Institution's needs can be investigated, and complementary solutions can be developed.

Definitions

Though SI will need to formally define research data, the pilot team's working definition is based on the U.S. Code of Federal Regulations[6]: "Any information used as primary resources to support research or scholarship...used as evidence in the research process and/or [that] are commonly accepted in the research community as necessary to validate research findings and results. All digital and analog content have the potential of becoming research data."

And thus, research data *management* encompasses all the activities surrounding data throughout the research process including planning, collecting, organizing and analyzing, describing, storing, sharing, and preserving.

> Good data management is not a goal in itself, but rather is a key conduit leading to knowledge discovery and innovation.[7]

[6] Electronic Code of Federal Regulations Sub A Chap 2 pt 200 https://www.ecfr.gov/cgi-bin/text-idx?node=2:1.1.2.2.1&rgn=div5#se2.1.200_1315

[7] Wilkinson, M. D. et al. 2016

RECOMMENDATIONS FOR IMPLEMENTATION OF A RESEARCH DATA MANAGEMENT PROGRAM AT SI

To begin to programmatically address the management of data assets, the Institution must begin to think as One Smithsonian and look beyond the short-term requirements for specific projects, researchers, and publications. Following successful examples both within and external to SI, the pilot team recommends the Institution adopt a combination of centrally coordinated services, unit-embedded data experts, researcher training, and increased support from pan-Institutional units including OCIO, SIA, and SIL.

> Museums like the Smithsonian are one of the original big data projects. They have collected and preserved information for centuries through artifacts, specimens, and archival documents. Museums have initiatives in bioinformatics and data accessibility, but—so far—results are scattered and inconsistent. We have a long way to go.[8]

In order to do this, SI will first need to create a central program office to coordinate services, develop and enforce policies, and foster collaboration and communication among the various staff and units involved in creating and managing research data. Policies and support services will need to be developed with cooperation from all parts of the Institution – research units, organizational support units, and administration. Technical infrastructure, including a locally managed data repository and robust storage capacity, will need to be strengthened.

The pilot team strongly recommends the Institution review, debate, and adopt the following four high-level recommendations, as well as immediate next steps, to move the Smithsonian towards a sustainable approach to managing the Institution's research assets.

1. DEFINE AND INSTITUTE POLICIES THAT SUPPORT *FAIR*[9] DATA PRINCIPLES (APPENDIX A)

As a first step, the Institution must develop clear definitions for digital research data and craft policies that support SI goals, including adopting a "digital first" strategy and "driving visionary, interdisciplinary research and scholarly projects"[10]. One way to do this is to create policies that promote data sharing and support FAIR data principles, that is, creating data that is Findable, Accessible, Interoperable, and Reusable[11]. As suggested in the two previous studies, the pilot team believes it would be most effective to create a Smithsonian Directive (SD) for Digital Research Data that clearly defines roles and responsibilities for administrators, researchers, information professionals, and staff in support units.
In order to create this SD and develop other necessary guidelines and practices a Policy Working Group, consisting of senior researchers and leadership from concerned units, should be formed as soon as possible. A draft charter for such a group is provided in Appendix A.1 of this report.

A new SD, or related guidelines, should include policies that support FAIR data practices such as:

---

[8] Rogers, J. Daniel and Wendy Cegielski. *Convergence may help scientists predict the future*. Smithsonian Magazine. March 2018. Accessed 4/16/2018 http://www.smithsonianmag.com/blogs/national-museum-of-natural-history/2018/03/01/convergence-may-help-scientists-predict-future/

[9] Findable, Accessible, Interoperable and Reusable https://www.force11.org/fairprinciples

[10] Smithsonian Institution. 2017. Goals 3 and 5

[11] FORCE11. 2014

- Mandating Data Management Plans (DMPs) for all research projects that create data.
- Set standards for minimum acceptable descriptive information (metadata) for datasets, including use of globally unique identifiers (GUIDs.)
- Providing guidance on preferred sharing platforms and repositories.
- Promoting sharing data under open licenses or with minimal restrictions, as appropriate.
- Providing guidance on appraisal and/or retention criteria for research data.

In order to support those policies, the Institution will need to create incentives for researchers to openly share their data and provide resources to enable them to meet policy standards and goals. Incentives could include internal grant funds for special projects or direct assistance with data description and analysis. Like the publication of peer reviewed articles, or curation of exhibitions, encouraging and rewarding the sharing of research data acknowledges the central role data plays in powering large sections of the world economy as well as driving innovation and knowledge creation. As such, the Institution should:

- Ensure data publication and sharing is a measurable and creditable research outcome in PAEC[12] reviews.
- Provide a central registry of datasets to track their publication and re-use.
- Provide staff (data managers) and tools to support creation of necessary documentation (DMPs and descriptive data) as well as perform quality assurance on 'submitted' datasets.

2. ESTABLISH A CENTRAL RESEARCH DATA MANAGEMENT PROGRAM OFFICE (APPENDIX B)

The Institution should create a central RDM Program Office to create efficiencies, set priorities, and coordinate common needs like training, metrics, and systems management. This office can work with units to advocate for the larger goals of the Institution with regards to research data and foster communication among researchers, and between researchers and information professionals. Ideally the central Program Office should be located outside any one unit reporting structure to emphasize its pan-Institutional intentions and neutral point of view. The work of the Program Office should be complemented by subject matter experts with data skills embedded in the units, and expanded services from pan-Institutional units already engaged in RDM and digital preservation initiatives. To combine these efforts and create a pan-Institutional RDM network, SI should:

- Create a central Program Office that reports either to the Provost or the DUSCIS[13].
- Hire data management specialists with subject matter expertise that are embedded in research units.
- Distribute supporting roles to and build upon existing strengths in pan-SI units:
  - OCIO – technical infrastructure support, collectively needed programming and data science services.
  - OSP – referral to appropriate resources and units for proposal technical needs (DMP, data modeling.)
  - SIA – consultation on format standards, assessment of data for permanent retention, and digital preservation activities.
  - SIL – consultation on standards for description, registering citations for datasets (metrics), and training and education in RDM best practices.
- Work with external organizations to promote and improve data sharing and management capacity.

Following the successful model of the Smithsonian's National Collections Program, the pilot team also recommends that seed funds be identified to create a "pool" fund from which internal grants can be awarded to address special projects or create additional capacity for RDM in a unit.

[12] Professional Accomplishment Evaluation Committee, or Professional Accomplishment Evaluation Panel
[13] Deputy Under Secretary for Collections and Interdisciplinary Studies

3. PROVIDE COOPERATIVE SERVICES THAT SUPPORT *FAIR* DATA PRACTICES (APPENDIX C)

In order to effectively manage research data a number of support services will need to be provided to researchers by both a central program office and pan-Institutional units. Some of these services, such as administration of a central data repository, are already in place, and should be expanded or refined before more aggressively promoting them to the SI research community.

Effective data management processes, though, begin even before data are collected or created. Planning data projects, doing data modelling and analysis, and creating quality data description are a distinct set of skills on top of those mastered to do research in the sciences or humanities. As such, it is imperative that individual training and support be available to everyone who creates, collects, or works with data.

Based on surveys of current researchers and investigations into similar suites of services provided at comparable institutions, the pilot team recommends that priority be given to the following services, some of which can be delivered using existing resources in the Libraries and OCIO. To offer these services to the larger SI community, however, SIL and OCIO should formally devote more resources in the form of staff time or funds to these activities. Essential activities include:

- Providing both in person and online training in RDM related topics for new hires, interns and fellows, and periodically for any staff who work with or create data.
- Engaging in communication and community building, including providing opportunities for peer-to-peer information sharing such as a yearly symposium or "share fair."
- Supporting shared RDM tools and platforms, including a new data repository, as well as commonly needed software such as Electronic Lab Notebooks.
- Assisting with selecting descriptive standards (metadata), creating and reviewing DMPs.
- Extending the capacity of units with no central IT or web office, e.g., by helping to create SOW for custom programming, and through providing centralized data science services.
- Supporting delivery of metrics for datasets by incorporating dataset citations to the existing Smithsonian Research Online (SRO.)

After a program office has been created and additional staff have been hired, the Institution must provide services in support of the first and later phases of the data lifecycle, specifically planning and preservation and curation such as:

- Quality review of datasets and descriptions prior to deposit in an SI repository.
- Special project to describe, prepare, and deposit legacy datasets, particularly those on outdated media.
- Ongoing curation, management, and migration of datasets.

The Smithsonian Archives will be a key partner in developing the policies and workflows for preservation and curation of research data, and integration of RDM into a larger digital preservation program at the Institution.

4. STRENGTHEN TECHNICAL INFRASTRUCTURE (APPENDIX D)

Researchers' capacity to produce data is only accelerating. In order to plan for future needs, the Institution should begin any evaluation of its technical infrastructure by conducting a formal inventory of research data at SI. A review of SI's technical infrastructure around research data comes as the Institution begins to implement a new five year strategic plan.

There are three major components of a robust technical infrastructure to support RDM – adequate network throughput and pipelines for moving large files, storage, and file and metadata management. To support sharing

and archiving across a wide range of projects, and to fulfil part of SI's new strategic goal to "create new digital platforms for scholars and educators to better access Smithsonian collections, research, and education resources[14]" SI must be able to provide several options for deposit of data. For researchers who are not directed by funder or project mandates to deposit in a particular external repository, SI should provide a locally managed alternative that provides functionality needed by the majority of projects.

Currently SI manages two local repositories that can be used for datasets, SIdora and SRO. SIdora is the most versatile of the SI repositories, with an ability to handle a wide variety of file formats and large datasets. However, lack of wide adoption by researchers, along with pending platform obsolescence, leads the pilot team to recommend replacing and/or supplementing SIdora with an equally robust but more user-friendly platform, followed by a re-evaluation in three years. Any repository platform SI adopts should also be evaluated for its ability to support digital preservation activities**.**
The following activities will help SI define the target e-research infrastructure at the Institution:

- Conduct a formal assessment of existing research data volume.
- Incorporate RDM technology needs into any technology planning, specifically large file transfer and storage.
- Supplement existing SIdora data repository with a new system(s).
- Work with SIA to explore systems and technologies that enable digital preservation.

NEXT STEPS

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

- 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.
- Establish a central Research Data Management Program Office, modeled on the Smithsonian's National Collections Program.
- Identify resources to staff the Program Office and data manager positions, and seed a RDM "pool" fund.
- Quantify at-risk research data, and facilitate infrastructure planning, by conducting a formal inventory.

MAINTAIN MOMENTUM

Until the recommendations outlined above can be fully implemented, the units represented by the pilot team – OCIO, SIA, and SIL – will need to continue to move the Institution towards programmatic research data management. SIL in particular should dedicate additional staff time to providing training and outreach for RDM, and both SIL and SIA should set aside staff time to work with OCIO to implement a new data repository.

- [SIL] Expand training opportunities, including cross-training, for research and other staff in RDM.
- [SIL] Begin collecting dataset citations in SRO to provide metrics to demonstrate research impact.
- [RDMPP team[15]] Plan early CY19 RDM share fair, identifying any additional funds needed.
- [RDMPP team] Participate in the Digital Library Federation eResearch Network[16] (May-October, 2018.)
- [OCIO] Complete repository testing against requirements and set timeline for launch.
- [OCIO] Include RDM infrastructure needs into any future technology planning, e.g., pipelines for very large file transfer and storage.

[14] Smithsonian Institution. 2017
[15] With assistance from the SIL Data Management Working Group, which includes staff from around the Institution.
[16] https://www.diglib.org/opportunities/e-research-network/

APPENDIX A: RECOMMENDATIONS FOR RESEARCH DATA POLICIES

Method

To evaluate what policies are necessary to facilitate the management of digital research data, the pilot team started by reviewing recommendations from earlier studies and looking at existing Smithsonian Directives (SD) and policies related to research and digital assets including the Public Access Requirement for Federally Funded Research ("Public Access Requirement"). The pilot team also analyzed policies from Federal agencies and select universities, and reviewed general best practices publications related to policy development and implementation.

> Research institutions should establish clear policies regarding the management of and access to research data and ensure that these policies are communicated to researchers. Institutional policies should cover the mutual responsibilities of researchers and the institution in cases in which access to data is requested or demanded by outside organizations or individuals.[17]

Overview

Though there is significant overlap in SI policies related to research data, there is also a significant gap in guidelines for managing active data, raw data, or data (in SD 501 referred to as "records") that are of temporary or long-term, but not permanent, value. In practice active data are managed at the discretion of the individual researcher. However, decisions made in the early phases of the data life-cycle, such as which descriptive standards and formats are used, can have significant effects on the ability to properly share and preserve those data later. As such, and in agreement with the recommendations in both the 2016 AVPreserve report[18] and the 2011 OPandA report[19], the Institution should develop a Smithsonian Directive specifically for digital research data.

The pilot team recommends that a Research Data Policy Working Group, made up of senior researchers, unit directors or their designates, and qualified representatives from support units be chartered as soon as possible (see Appendix A.1). This group – or possibly groups – will work to investigate and understand the major issues that should be addressed by policies in order to effectively and programmatically manage research data at the Institution, including defining what constitutes "research data" at SI. They will evaluate the existing SDs and policies that apply to research data, and draft an SD for digital research data, along with any necessary related guidelines or technical notes, and if necessary propose amendments to existing policies to harmonize the Institution's guidelines for research data.

> Finally, only by sharing research data and the results of research can new knowledge be transformed into socially beneficial goods and services.[20]

Summary

- SI should create an SD for digital research data.
- The SD and related policies should include practices that support FAIR data principles including:

[17] National Academy of Sciences. 2009. Recommendation 8
[18] AVPreserve. 2016 Recommendation 6.4.3, p. 79
[19] OPandA. 2011 Recommendation 4, p. xvii
[20] https://www.ncbi.nlm.nih.gov/books/NBK215271/#ddd00054 sect 3

    - Mandatory DMPs that include essential information for sharing, preservation, and re-use of data.
    - Open licenses (where appropriate) for re-use.
    - Use of GUIDs and sufficient metadata to enable citation and discovery.
- Give researchers incentives for publishing data such as:
    - Ensuring data is a measurable and creditable research outcome in PAEC[21] reviews.
    - Providing metrics to show use and re-use (see Appendix C.)
    - Providing staff and tools to support creation of documentation as well as perform quality assurance on published datasets (see Appendix C.)

Existing policies

SI has three existing directives, *SD609 Digital Asset Access and Use*[22], *SD610 Digitization and Digital Asset Management Policy*[23], and *SD501 Archives and Records of the Smithsonian Institution*[24], that include provisions for the management of Institutional digital assets, which can include research data. SD 609 deals exclusively with policies for sharing data, and applicable exclusions for sharing. SD 610 covers the creation and management of digital assets, explicitly including research datasets that are "created, stored, or maintained by SI."[25] SD 610 does not, however, explicitly cover datasets produced through collaborative projects, or those which are not permanently stored at SI, e.g., genomic sequences in GenBank. Asset lifecycle policies, such as disposition and archiving schedules, are the responsibility of the units and are to be detailed in digital asset management plans[26] (DAMPs). Responsibilities for carrying out the details of any unit's DAMP, however, are not explicitly included, and in practice the Institution has not been able to ensure that all data producers also create DAMPs.

SD 501 specifies the responsibilities of SI employees and SI Archives for handling "All documents created or received by employees of the Smithsonian Institution in the course of official business"[27] and records of permanent value that preserve the history of research activity at SI. SIA defines research records in its official disposition schedule[28] as:

> …any materials created or collected while conducting research as part of an employee's official duties. They may exist in many formats including sheets of paper, notecards, field books, maps, databases, spreadsheets, audiovisual materials, and images. Research records can typically be divided into three broad categories: 1) secondary research materials; 2) original research materials; and 3) research outputs. In each of these categories, there may be a subset of materials often referred to as 'working files.'

Though SIA provides extensive guidelines on management of active files, they are not responsible for management of those research assets until the data becomes inactive and is transferred to SIA according to pre-existing disposition schedules, or in consultation with the Archives. There are no specific mandates, only guidelines, for research owners on how to manage their data.

---

[21] Professional Accomplishment Evaluation Committee, or Professional Accomplishment Evaluation Panel
[22] http://prism2.si.edu/SIOrganization/OCFO/OPMB/SD/SD609.pdf
[23] http://prism2.si.edu/SIOrganization/OCFO/OPMB/SD/SD610.pdf
[24] http://prism2.si.edu/SIOrganization/OCFO/OPMB/SD/SD501.pdf
[25] SD610 part V, p. 4
[26] SD610 part VIII, p. 9
[27] SD501
[28] https://siarchives.si.edu/what-we-do/records-management-faq

The Public Access Requirement only covers data related to a peer-reviewed publication, not a larger set of underlying raw data or data that is unrelated to an existing or forthcoming publication. It also is specifically focused on the access to that published data through sharing, and not on other phases of the data lifecycle, such as preservation.

Though there is overlap in all of the existing policies, in practice there is a gap into which many active datasets and long-term research studies may fall. There is no explicit coverage for data produced in collaborative projects that may have poorly-defined ownership. There are no guidelines for how and when to share data that are not part of a publication. There are also no incentives for data owners to follow best practices and guidelines for managing their data, and no consequences other than implied loss. Therefore, any new policies must address all the above issues, as well as roles and responsibilities for managing data while it's being collected and analyzed, when sharing data, as well as providing guidance for responsibly stewarding data for the long-term once a project or dataset is finalized.

Research data policy coverage[29]

Because Smithsonian Directives have a standard structure and common features, even if an SD is created for digital research data, it may be necessary to also draft supplemental "technical notes" or incorporate additional information into other SDs or policies, particularly SI's Public Access Requirement. The following outline includes the broad topics that would be expected in RDM policies, all of which should be included in some SI policy or guideline. An outline of a policy, with some suggested topics and content can be found on the RDMPP's Confluence site.[30]

- Scope and definition
  - Definition of research data
  - Scope of policy vs. other SI policies
- Ownership, rights, and restrictions
  - Intellectual Property Rights and other allowable restrictions
  - Default usage rights and licensing, or how to determine rights
- RDM standards
  - FAIR principles and description (metadata) standards
  - Use of unique identifiers
  - Creation and essential content for DMPs
  - Retention periods, assessment process, and disposition[31]
  - Preferred repositories
- Responsibilities
  - For researchers, supporting units, and SI broadly including:
    - MOUs and other agreements for collaborations
    - Costs, e.g., for storage
    - Data quality
    - Archiving and future curation

In general, any policy around digital research data at SI should support the FAIR data principles. In order to be FAIR, the pilot team recommends any policy include minimum descriptive standards (including use of GUIDs),

[29] Adapted from Leaders Activating Research Networks (LEARN). 2017. Guidance for Developing a Research Data Management (RDM) Policy http://learn-rdm.eu/en/learn-toolkit-download-individual-sections/

[30] https://si-confluence.si.edu/display/RDMPP/Guidelines+for+Managing+Digital+Research+Data+at+SI

[31] This might include how to determine retention periods for parts of datasets, e.g., raw data vs. analyzed data, for particular fields of study.

mandatory DMPs, best practices for choosing file types and formatting, and use of preferred repositories that meet FAIR standards. As one of their deliverables, the pilot team produced several best practices documents on data management that cover those topics and more. They are currently available on the SI Libraries' website.[32]

Including descriptive information (metadata) with data is crucial to make them findable and citable. Use of GUIDs, such as DOIs, offer a stable and simple way to cite data, and can facilitate gathering additional use data which can benefit data producers when it comes time to demonstrate research output. Providing detailed metadata in a "data dictionary[33]" is also essential to give future researchers sufficient information to reconstruct the context surrounding the dataset – how it was generated and analyzed, what standards were used – that will enable them to re-use the data in a meaningful way or reproduce the results of the original study.

DMPs, while ostensibly existing to plan a research project, provide essential information at the completion of a project. A properly written plan can transmit the intentions of the data creators to repository managers and data curators as to how the data should be maintained and shared, as well as providing context to enable any necessary transformation or format migrations. Though DMPs are required by some funders as part of a grant application, the mandatory elements of those plans are often insufficient to provide all the information future curators may need to care for the data. In order to begin to "future proof" SI assets, the Institution should require DMPs for all research projects that produce digital data.[34] An added benefit of this requirement, if DMPs are provided to and reviewed by OCIO, is that it gives OCIO advanced notice when planning for and acquiring additional storage capacity. In conversation with SI researchers, two staff PIs independently recommended that the Institution require DMPs. In order to facilitate this, however, any RDM program will need to provide training and assistance in creating plans, as well as a review process to ensure they are complete.

> We found that most responding agencies agreed that requiring research proposals from intramural and extramural scientists to include data management plans is an important component of their plans to comply with the memo. [35]

The Smithsonian should also decide if SI scholars should be directed to use a particular set of data repositories, and if so what they are and what criteria is used to select them. A recommended set of repositories may be chosen for legal reasons such as support of particular licensing schemes, for philosophical reasons like promoting Open Access, or for practical reasons like tracking and monitoring research assets.

Incentives

Besides just creating mandates on how to manage data, the Institution must provide incentives for researchers to comply with those mandates, as well as support for complying. There are several ways this could be accomplished. The most basic is to ensure that published data that conforms to SI policies is included as a creditable research outcome in all PAEC review activities.

---

[32] https://library.si.edu/research/data-management

[33] A data dictionary, sometimes called a readme file, or just metadata, is a file that describes each element in a dataset – for instance, it might include a list of the fields in a tabular dataset and what they mean, including units and precision. For a brief guide see http://datadryad.org/pages/readme

[34] How to apply this will depend on how the Policy Working Group defines "research data" and how the program is staffed.

[35] Kriesberg, A. et al., (2017). *An Analysis of Federal Policy on Public Access to Scientific Research Data*. Data Science Journal. 16, p.27. DOI: http://doi.org/10.5334/dsj-2017-027

Formally recognise and value data, alongside publications, as part of research assessment and career advancement[36]

Units should also formalize adherence to policies by incorporating them into researcher performance plans. In order to provide documentation for data publication and re-use, the SI Libraries will need to expand their collection of citations in SRO to include citations for datasets (see also Appendix C).

The barriers to compliance with policies must also be lowered through providing services such as training and DMP creation assistance and review (see also Appendix C). Most importantly, though, SI must provide adequate technical infrastructure to enable good RDM practices including tools and software for managing data, adequate storage, and a repository that encourages creation of essential metadata while providing a user-friendly interface that also gives researchers control over how their data is seen and shared.

As a premium incentive, and following similar programs at the Institution (see Appendix B), the RDM program should provide "pool" funds for which units, labs, or researchers can apply to further their projects. These funds, similar to those managed by the National Collections Program, would be competitively awarded and would require adherence to best practices and policies as a primary condition for acceptance.

[36] Van den Eynden. 2014

APPENDIX A.1 RESEARCH DATA POLICY WORKING GROUP

Charter

SI Research Data Policy Working Group

Version 0.11

Background

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 the American people to steward these assets responsibly and share them broadly. Managing digital research data in a way that enables future access and re-use serves not only to fulfil the mission for the "Increase and Diffusion of Knowledge" but can strengthen the public trust in the scholarly process and its outcomes by promoting reproducible research and verifiable facts.

In 2016[37] and 2011[38] two reports related to research data management (RDM) were published. In each, they recommended that SI develop policies for data management and sharing. The 2016 report specifically recommends creating a Smithsonian Directive for management of research outputs. In addition, many funders, including the National Science Foundation, require that grant proposals include a formal Data Management Plan that contains documentation of institutional policies around data.

Without a strategy and policies to guide progress and promote systematic data management, the Institution will remain mired in "small-science, seat-of-the-pants, fragmented, and mostly unit-or department-based[39]" efforts.

Purpose

This WG shall investigate and understand the major issues that should be addressed by policies related to effectively and programmatically managing research data at the Institution, including defining what constitutes "research data." The objective of the working group is to determine what policies are necessary for SI, and produce draft policies and guidelines that will be adopted at the Institution.

As necessary, the Working Group should break into sub-groups to accomplish this objective.

Working Group Activities

1. Formally define research data for SI.
    a. Scope the definition, as distinct from other types of data at SI such as collections data, publications, or ancillary data.
    b. Determine if research data is a product of staff working only in research units, or can be the product of any unit. Similarly define which type of staff, e.g., permanent, fellows, or visiting researchers, are affected by policies developed by the working group.

[37] AVPreserve. 2016. p. 180
[38] OPandA.2011 Recommendations p. 122
[39] OPandA.2011. p.xiii

2. Determine if policies for data management and sharing should be integrated into an existing Smithsonian Directive (SD) such as SD 503 or SD 610.
    a. If not, draft an SD for digital research data, and begin the approval process.
3. If the WG decides to create a new SD, draft interim guidelines for RDM that identify
    a. Roles and responsibilities for management of data throughout the life-cycle.
    b. Acceptable and/or preferred platforms for storing and/or sharing data, including roles for existing repositories at SI.
    c. Acceptable or preferred extent of data description (metadata) to facilitate discovery, reuse, and preservation.
    d. Mandatory activities and best practices for data management (DMPs).
    e. Policies for sharing, including preferred licensing schemes.
4. Update the Public Access to Federally Funded Research Requirements to conform to the new guidelines.

Membership

Members actively participate as equal and contributing technical and/or business resources. Members are expected to attend all or at least 80% of the meetings prepared, having performed all background reading and ready to offer constructive comments and suggestions.

Suggested Members

1. (OUSMRP/DUSCIS)
2. (OCIO)
3. (SIL)
4. (SIA)
5. (SOAR)
6. (OGC)
7. (SCBI)
8. (SERC)
9. (SAO)
10. (NMNH)
11. (MCI)
12. (NMAH)
13. (NMAAHC)
14. (SAAM)
15. (HMSG)
16. (FSG)

Deliverables

Draft policies and guidelines for RDM.

Timeline

APPENDIX B: RECOMMENDATIONS FOR PROGRAM STRUCTURE AND ADMINISTRATION

Method

There have been numerous publications[40] in the Library world that tackle how to develop a RDM program, primarily for Universities. The pilot team began with a review of those, and the existing programs at several Universities in the U.S., U.K. and Australia. Federal agencies have also developed RDM programs, some quite robust, in support of their missions and to comply with the 2013 White House OSTP memo[41] encouraging agencies to provide access to all Federally funded research. The pilot team interviewed staff from programs at NOAA and NAL/USDA, and spoke informally with staff from USGS. To supplement those models, the team also turned inward to look at successful pan-Smithsonian programs that combined one or more of policy, technical, and advisory services – these include the National Collections Program (NCP), the Digitization Program Office (DPO) and Smithsonian Research Online (SRO). The pilot team also returned to the two previous studies for related recommendations.

Overview

Program structure and staffing levels will be affected by any new policies adopted by the Institution and the corresponding services needed to enable compliance with those policies. In the absence of new policies, recommendations will focus on the essential resources needed for a program to start, and provide a picture of what a fully developed program might look like.

The Smithsonian has several operational gaps that need to be bridged in order to begin to systematically manage research assets. The primary gap should come as no surprise to anyone familiar with large organizations, that is, the inherent difficulties in communicating, and therefore coordinating efforts, within and across units. Though some units or labs may have researchers versed in RDM best practices, they may not practically have enough time to fully apply that knowledge, or share their expertise with others. There is also a lack of IT resources, in the form of programming or other engineering staff, in many of the research units to support their technical needs for data collection and sharing. When research data policies are established, the Institution will need resources to implement them, including staff to monitor policy compliance and measure success.

There are many possible ways to structure a RDM program at SI. The pilot team created four models[42] to use as a starting point for discussion by decision makers. All models are designed to provide a majority of the necessary services for a successful program including support for planning; policy development and compliance; education; data modeling and description; sharing and storing; and finally curating and preserving.
The recommended program structure, however, considers more than just how to provide support services. A successful pan-Institutional program at SI must create a sense of community among researchers, information professionals, and collections staff – bringing them together to share what works, building capacity for the entire Institution. It should also leverage existing work and natural strengths in the non-research units related to RDM, particularly OCIO, OSP, SIA, and SIL.

---

[40] Jones et al. 2013, McGinty et al. 2012, Fearon, et al. 2013, Bryan et al 2018.

[41] Executive Office of the President. Office of Science and Technology Policy. 2013. *Memorandum: Increasing Access to the Results of Federally Funded Scientific Research*.
https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf

[42] Specific characteristics of each model, as determined by the pilot team, are available on the RDMPP Confluence site https://si-confluence.si.edu/display/RDMPP/Program+structure

Summary

- Create a central program office to build community and facilitate communications, manage service coordination, policy development and compliance.
- Hire domain experts with data skills to facilitate RDM and embed them in research units.
- Distribute appropriate supporting roles to existing pan-SI units:
  - OCIO – infrastructure support, including collectively needed programming and data science services.
  - OSP – referral to appropriate resources and units for proposal technical needs.
  - SIA – assessment of data for permanent retention, curation and preservation activities.
  - SIL – assistance with metadata and standards for description, registering DOIs and citations for datasets to provide metrics, training in RDM best-practices.
- Work with external organizations to promote and improve data sharing and management capacity.

Models

Federal Agencies tend to emphasize technical infrastructure at the center, and either a top-down compliance model, or a combined top-down and broadly cooperative model for program organization. The pilot team looked at three agencies – USGS, NOAA, and USDA/NAL, chosen because they had established programs and created data that in some cases was similar to those created at SI. Federal exemplars, however, have three key differences from SI - they have technologically and resource intensive programs limited to science; they produce large quantities of similar data, e.g., geographic surveys; and they award grant funds both inside and outside the agency which provides leverage for policy compliance. Because of these key differences, the team does not recommend that SI adopt their program structures, though we do encourage SI to consider their successful communication strategies and policies as models.

Universities, which mirror SI in terms of the heterogeneity of research output and the independence of both researchers and departments (units), increasingly provide RDM services through the library. Because of the (on average) small size of these library-run programs, there is often less emphasis on technical infrastructure and more on outreach and training. Depending on the specific university, varying amounts of data management work may be done at the department level or from within, or in cooperation with centralized IT. Unlike SI, many universities have a central Dean of Research (or similar) that sets policies and provides incentives and oversight.

Looking inward, there are several pan-Institutional programs that support some combination, or all, of the classes of services in RDM such as technical infrastructure; planning and policy development and oversight; and/or training and communication management. We considered three existing programs as models. The National Collections Program (NCP) provides high level central planning and policy development, centralized reporting, and limited direct services to units, with technical infrastructure decentralized and supported through OCIO and unit staff. The Digitization Program Office (DPO) provides high level central planning and policy development, centralized reporting, some centralized and shared technology (such as for 3-D digitization), select training opportunities, and some direct services (through special projects such as the Botany and Paleobiology department mass-scanning). Smithsonian Research Online (SRO) includes centralized technology, limited policy development, centralized reporting, and centralized services in the form of mediated deposit and citation harvesting.

The Research Data Management Program

The pilot team considers the "NCP model" to have the most potential for successful implementation because it combines a lean centralized program office with a neutral pan-Institutional perspective and experts distributed in the units. It also has an incentives component in the form of competitive grants that require planning and metrics documentation.

RDM, like collections management, requires multiple communities of practice with specialized expertise who are best located on the "front lines" in the units. Conceivably, permanent staffing for this model could be phased in, beginning with two program officers and a cohort of four or more centrally funded and administered term fellows or postdocs serving as data managers in the units. The data managers would be subject area experts[43], possibly post-Docs, who understand the domain of the unit in which they are embedded. Their role would be to connect the work done in the unit with the program office and other supporting pan-Institutional groups as well as work directly with researchers and to institute best practices including creating data management plans and developing workflows for data collection and analysis that include crafting data descriptions (metadata).

At the far end of the data lifecycle, SI will need to have digital preservation and data curation experts to manage archiving and preserving research data deposited with the Institution. Ideally, a minimum of two data curators would be hired within the first two years of the start of the program. Reporting either to the central program office or embedded within the SI Archives reporting structure, they would be responsible for working with the Archives to develop an appraisal and deposit process for datasets, as well as perform Quality Assurance on data that is to be stored and shared in a local data repository. Their expertise will become invaluable when digital preservation projects come online to deal with SI's legacy datasets and research on outdated media.

A central program office, however, is still needed to foster communication and collaboration, create efficiencies, set priorities, coordinate common needs (training, metrics, systems) and advocate for the larger goals of the Institution. Like NCP, the central program office should ideally be located outside any single unit reporting structure, i.e., reporting directly to the Deputy Under Secretary for Interdisciplinary Support and Collections or to the Provost, to emphasize its pan-Institutional intentions and neutral point of view. Being administratively situated in a more central position in the Institution would also provide a certain amount of weight when working with unit administration.

The program office initially would be responsible for shepherding the policy development process and later would transition to managing compliance with policies. They would connect data producers, resources, and support staff through activities like organizing yearly gatherings ("share fair" or symposia) where research staff can share information and learn about pan-Institutional resources and initiatives. Periodic (possibly quarterly) peripatetic open houses or face to face meetings with researchers would also be valuable. The program office would also work with OCIO to administer the local data repository. When identified, the program office would also be responsible for administering competitive grant funding, similar to the funds available from NCP. In RDM, these funds could be used to target projects on specific at-risk data or develop tools that can be shared across units.

When fully staffed the office could work on special projects or for resource-intensive priorities, such as planning and assisting with legacy data description for "orphaned" data whose creators are long gone. They would also be able to devote resources to participating in RDM initiatives beyond the Smithsonian (see below) and working with the Office of Fellowships and Internships to develop partnerships with universities to provide real world training for those interested in data management and data curation.

The following staffing scenarios use GS levels[44] and salaries to demonstrate the potential cost and levels of expertise needed in each position. This is not meant to imply that all positions must be permanent and/or Federal, particularly at the start of the program. It may be beneficial to start out with some of the data scientist and data manager positions as temporary term or fellowship positions designed for pre- or post-Docs, to give time for units and central SI to gain a better understanding of the workload and of skillsets needed, and identify

[43] Effective data management in highly specialized fields, particularly at the beginning of a project when planning and collecting data, requires staff with both domain expertise and IT/data skills. It is often more practical to train domain experts in RDM, than train information professionals in very specific scientific fields.

[44] GS levels and position titles from http://www.fedsdatacenter.com accessed 2018-02-21

funds for permanent positions. Note the scenarios do not include resources that may be needed from other units, such as library liaisons, or potential contractor positions, e.g., in research computing or application development.

Start-up staffing

10 FTE staff at approximately $866,000 for 1 year

- 4 FTE data managers – embedded in the units GS11 - $300k
- 2 FTE data scientists – attached to OCIO GS13[45] - $214
- 2 FTE data curators – attached to SIA GS11 - $150
- 1 program manager GS14 - $127
- 1 program coordinator GS11 - $75

Fully staffed program

19 FTE staff at approximately $1.6 million for 1 year

- 8 data managers embedded in the units GS11 $600k
- 4 data scientists—divided between OCIO and units GS13 $428k
- 4 data curators – attached to SIA GS 11 $300k
- 1 program manager GS14 $127k
- 1 program coordinator GS11 $75k
- 1 program assistant or metadata librarian $62

The RDM Team – more than a program office

Support from pan-Institutional units will be crucial to the success of any program, particularly in the early stages of development. Even when a Program Office is fully staffed, it will not have the variety of specialized knowledge and skills that already exist in the units.

Cooperation from the Office of Sponsored Projects (OSP) is necessary to give the future Program Office notice of pending grant-funded projects so that they can become involved in the planning process. Ideally, the Program Office would be involved before a proposal is submitted to OSP, but where they are not, OSP would be a critical communications link. To ensure that the Program Office and other unit staff are adequately supporting grant proposal development and project planning needs, OSP should include notification of all new proposals to the Program Office so that they can contact the PI and request a copy of their DMP.

Besides providing technical support of infrastructure such as high performance computing (HPC), storage, and server management, OCIO is a critical source of IT expertise and institutional knowledge (see also Appendix D.) The Office of Research Computing runs the existing data repository, SIdora, and is intimately familiar with the research data needs of many at the Institution. As such, they should continue to run any future data repositories, but with a primarily technical rather than an administrative focus. Administration of the repository can be handled by the Program Office, freeing up Research Computing staff to focus on supporting the IT-specific elements of project planning, and coordinating work with other departments in OCIO to provide shared contracts for programming or development of tools and pipelines. Research Computing should also continue to provide the existing data scientist fellowships that hire experts to work on large-scale data analysis needed by projects that use HPC.

[45] This is below the 2016 industry average salary for the mid-Atlantic – see King, John and Roger Magoulas. 2016.*2016 Data Science Salary Survey*. O'Reilly Media, Inc.

The Smithsonian Libraries should, like OCIO, have a central role in the overall RDM program and work closely with the program office, unit data managers, and researchers. SIL's existing Scholarly Communications unit, which runs SI's publication repository and issues DOIs for datasets, is a natural choice to manage a "registry" of datasets. These citations for datasets deposited outside SI, like existing publication citations in SRO, can be used to create metrics on data use and reuse, as well as help the overall RDM program track the location of SI assets. SIL's research services staff, who are often co-located with researchers, are ideally situated to provide training in RDM best practices and coordinate work with other parts of the RDM program team, as well as assist with DMP preparation or data deposit. SIL also has several metadata and data specialists who would be available to work with researchers who need to select or develop descriptive standards. Academic and research libraries are increasingly playing an active role in RDM at their organizations. SIL is well-positioned to follow their examples, as was recommended in the Libraries' Data Management Working Group white paper from 2014. Ideally, when SIL chooses to truly commit to supporting data management at the Institution, besides ensuring Research Services staff have appropriate data-related skillsets, additional staffing for SRO should be provided to support quality assurance and review of datasets deposited with publications.

The Smithsonian Archives will be critical in developing and then implementing an appraisal process for data, and advising staff on retention and disposition of data. SIA can also work with the Libraries to develop and educate staff on best practices in selecting file formats and working with and organizing data. Prior to deposit, SIA staff will need to work with data managers in the units and central program office staff to provide quality assurance on datasets to insure completeness and policy compliance. When SI develops a technical infrastructure to support the ongoing preservation and curation of records after they are archived, the Archives will be responsible for managing those processes. As such, it may be in the Institution's best interests to provide additional funds directly to SIA to create additional positions and hire data curators, rather than have curators report to the central program office.

Funding

Once a program office is established, one of their first tasks should be to work with appropriate partners in OCIO and the SI Office of Planning, Management and Budget to prepare a business model for the RDM program that includes staffing, service, and infrastructure needs and provides for a small "pool" fund to support special projects and incentivize researchers to adopt best practices[46]. Several business models and funding sources (specifically for repositories, but applicable more broadly to RDM) along with pros and cons of each approach are outlined in the Organisation for Economic Cooperation and Development (OECD) *Business models for sustainable research data repositories*[47]. The conclusions in that paper point toward a diverse, or blended, funding approach to enable flexibility and avoid single failure points.

> Reliance on project funding for research data management…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.[48]

The AVPreserve report, though focusing specifically on digital preservation, echoes the recommendations in the 2011 OPandA report and the more recent 2017 OECD paper. In order to achieve sustainability, AVPreserve

[46] See Appendix A for more on the "pool" fund as an incentive.
[47] OECD. 2017
[48] OPandA. 2011 p.85

recommends a blended funding approach – a base of programmatic funding supplemented by a percentage of any project-based funding, e.g., through overhead charged to received grants.

The OPandA report makes several specific recommendations related to funding[49], including pursuing a new Federal line item just for data management; increasing and/or reallocating overhead allowances on grants; shifting funds from lower-priority functions; and leveraging external partnerships and other collaborative initiatives.

> Leveraging resources through partnerships and participation in collaborative initiatives is one important avenue to follow; Smithsonian data-management and sharing efforts are still too often undertaken in relative isolation from external organizations….The Smithsonian will need to pursue a combination of these and other strategies.[50]

Beyond the Smithsonian

Looking beyond the immediate resources available at the Institution, a future program could also benefit from collaborations with external organizations and efforts. Involvement in international groups like the Research Data Alliance[51] (RDA) as well as U.S.-based efforts like the Data Curation Network (DCN)[52] can provide material benefits to RDM at the Smithsonian. RDA is a community-driven organization that provides a "neutral space where its members can come together…to develop and adopt infrastructure [and standards] that promotes data-sharing and data-driven research, and accelerate the growth of a cohesive data community that integrates contributors across domain, research, national, geographical and generational boundaries." The Data Curation Network is a Sloan Foundation funded effort to develop a model for expertise sharing across institutions. Data curation and management staff at participating institutions (primarily U.S. universities) share their specialized knowledge as needed as well as promoting data curation training and best practices.

If resources are sufficient, the program could also support the creation of data curation capacity beyond the Smithsonian through developing and managing a data curation internship or fellowship program. Developed with assistance from the Office of Fellowships and Internships, the central program office could work with universities that have data curation certificates or similar programs to give students or recent graduates paid experience working with real data doing "data rescue" for staff or fellows who plan to leave SI within a year. Alternately, a Fellowship program could be developed to train pre- or post-Doctoral scientists and humanists in RDM practices. These Fellows could work alongside unit-embedded data managers, and rotate through underserved units to initiate new RDM projects.

[49] OPandA. 2011 p. xix
[50] OPandA. 2011. p. xv
[51] https://www.rd-alliance.org/
[52] https://sites.google.com/site/datacurationnetwork/

APPENDIX C: RECOMMENDATIONS FOR PROGRAM SERVICES

Method

As a starting point, the pilot team assumed that the services required for a comprehensive RDM program at SI would be similar to those used at other institutions. Several studies[53,54] have been published that explore and define those services for Higher Education Institutions. The pilot team supplemented those published studies with an independent review of established programs at several Federal Agencies[55] and data from informal interviews with and surveys of Smithsonian researchers.

Overview

The 2017 OCLC report on scoping the RDM service[56] divides services broadly into three categories – education, expertise, and curation – with increasing levels of resources needed for each category, curation being the most resource intensive. This is a similar model employed by the SIL Data Management Working Group in their 2014 report, and in the 2012 article by Reznik et al.[57] Within each of these categories, distinct activities around data can be supported at a level tailored to the institution. For example, metadata creation services – accurately and succinctly describing a dataset to enable discovery and reuse – could be addressed in at least three different ways. At a minimum, by providing group training on description standards for data writ broadly; providing dedicated staff who will consult with a department to create metadata templates for certain classes of data; or providing programming staff to create custom pipelines that process raw data to generate partially complete descriptive metadata files. In the context of an RDM program, each of those approaches requires very different staffing solutions. Policy decisions will also affect the services, and level of service, needed for any program. If SI adopts a policy that all data should be deposited in a local repository, providing repository services and adequate staffing to manage the repository will obviously need to be prioritized.

Summary

Any RDM program at SI should at minimum include:

- Training in RDM and related topics and tools both in person and online, for new hires and periodically for any staff who work with or create data.
- Communication and community building including providing opportunities for peer-to-peer information sharing such as a yearly symposium or "share fair."
- Support for project planning, DMP creation and review, as well as data modeling.
- Consulting on selecting descriptive standards (metadata) and best practices.
- Support for shared tools and platforms, including multiple repository options as well as commonly needed software such as Electronic Lab Notebooks (see more in Appendix D.)
- Extend the capacity of units with no central IT or web office, e.g., by helping to create SOW for custom programming, and by providing centralized data science services.
- Delivery of metrics for datasets, similar to those provided through Smithsonian Research Online for publications
- Quality review of datasets and descriptions prior to deposit in an SI repository.

---

[53] Jones, S., Pryor, G. & Whyte, A. (2013). 'How to Develop Research Data Management Services - a guide for HEIs'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

[54] ARL SPEC kit 334

[55] Specifically, the pilot looked at USGS, NOAA, and USDA/NAL

[56] Bryant, Rebecca et al. 2017.

[57] McGinty, Stephen et al. 2012.

- Ongoing curation, management, and migration of datasets (digital preservation.)

It is possible that some of these services – particularly those around the later part of the research data lifecycle such as curation – could be phased in after a program has been established and funding sources identified. Once a program is up and running, however, it should also support digital preservation activities around research data such as "data rescue" for at risk research data stored on outdated media.

Education and Training

Following on a survey done in support of their earlier report[58], in the summer of 2017 the Libraries' Data Management Working Group set up a table in the SCOS research tent during the staff picnic. They asked staff who dropped by to "vote" for specific services they thought would help them manage their data. The two surveys[59] returned similar results, with the most requested service as training, specifically workshops on best practices for managing data. Other highly ranked services included assistance documenting or describing data prior to deposit and assistance depositing data, including identifying an appropriate repository.

Because education and training were obviously desired, and because there were existing resources to address this need, one of the first things the pilot did was partner with staff in NMNH's L.A.B. to become a member of Data Carpentry[60]. Seven SI staff, including one pilot team member, are now trained as Data Carpentry instructors and four workshops[61] are scheduled for 2018. That same pilot team member, Sue Zwicker, is also providing RDM best practices training, with two classes already given at SERC and NMNH, and more planned for other units. All of these training activities have generated considerable "buzz" around RDM in the units that should be capitalized on.

Educational programs such as Data Carpentry workshops and best practices classes should be continued and extended to provide multiple channels for delivering training to staff on data description and organization, data manipulation and analysis, and related tools. Basic training in RDM principles should be provided for all new staff, interns, and fellows who do research as part of the on-boarding process. The Libraries (SIL) is a natural fit to provide these services, and should be involved in coordinating other data-related training as requested by staff.

Communication and community building

Following the example of many of the other programs that we investigated – particularly those Federal Agencies with widely geographically distributed staff – a central program office should foster communication and community building by providing periodic opportunities to gather in person and share experiences on what works in RDM. Other programs at SI, such as the NCP and DPO, also hold periodic "share fairs." These could be augmented by offering requested training on a particular topic to accompany a program of internal speakers and round tables. The program office could also develop monthly calls that include addressing a researcher's specific RDM need, as USGS does. The pilot team found that in-person interactions proved the most effective, however, and recommend that a program office consider holding quarterly web-cast meetings that move from unit to unit and not only provide information *to* unit staff in available services and resources, but also listen and gather information *from* unit staff on their projects and any friction they may be experiencing as they collect, analyze and share their data.

---

[58] SIL Data Management Working Group. May 2014. *Recommendations for Libraries' Involvement in Data Management at SI*

[59] A summary of both are available on the RDMPP Confluence site (LINK)

[60] http://www.datacarpentry.org/

[61] Workshops are focused on training in commonly used tools for working with data in Ecology and Genomics such as OpenRefine, Python, R, and SQL.

Planning and DMPs

Though many funding agencies require data management plans (DMPs) as part of a grant proposal, the templates for those plans are often brief and have open-ended questions which may not encourage the level of detail that would render those plans useful for researchers or data curators after a project's completion. The plan templates rarely include methodologies that should be followed to determine the contents of the plan itself. If SI adopts mandatory DMPs for data projects, as recommended in Appendix A, it will need to support this mandate by providing training and assistance to researchers in creating meaningful and actionable plans. This is a common type of training provided by academic libraries with RDM programs.

In order to intervene early enough in the planning stages of a research project, SIL staff should strive to be involved in all stages of the research lifecycle for staff in their unit, and provide regular reminders that DMP training and assistance are available. To catch projects that have gone through the planning stage and are in the proposal stage, the Office of Sponsored Projects (OSP) should provide a copy of any proposal that includes generation of data to SIL, OCIO, and central Program Office staff for review and comment.

After project approval or proposal success, DMPs should be kept for ultimate deposit with the data they apply to. This practice should eventually support the emerging concept of "Actionable DMPs[62]" which can be understood by machines and used to interpret data management needs at time of archiving. Including DMPs with their data can also provide additional context around the data, which reduces the need to create additional documentation, and can make re-use easier and more accurate. One option for creating and managing DMPs might be to use a shared online tool like the DMPTool.[63] DMPTool provides custom templates, with prompts and boilerplate[64] that can be customized to SI needs, and later exported as a PDF, MS Word document, comma delimited tabular file, or plain text file.

Technical support for tools and platforms

The pilot team conducted several interviews with select departments and labs at SI, mostly in SCBI, NZP, SERC, and MCI. These informal conversations turned up several additional service needs, most related to the lack of IT staff in each unit. Large collaborative projects, where several institutions may need to work together to gather and analyze data, are particularly in need of custom data collecting instruments, data modeling and database design, and custom pipelines for submitting both data and metadata from the field. Typically, sharing active data with staff on campus and in the field has been approached in low barrier ways such as emailing spreadsheets – or transferring large files by mailing hard drives – methods that don't scale and don't support standardizing metadata creation, e.g., through forms with drop-down menus. Static datasets, unless specifically related to a peer-reviewed publication, are often shared by putting them in cloud storage (such as Dropbox) or on a unit or lab website. While this has the advantage of being highly customizable, it has several disadvantages from a digital preservation aspect, and effectively silos published research data in the same way collections were siloed before the advent of Collections Search.

Though SI has a data repository, SIdora, that can be used to share as well as store research data, it is not particularly user-friendly and has not been widely adopted. SIdora's metadata requirements appear to be a barrier for some SI researchers. The pilot team strongly recommends the Institution supplement SIdora with additional repository platforms that can meet the need of a majority of data creators at SI (see Appendix D for

[62] Simms S, Jones S, Mietchen D, Miksa T (2017) *Machine-actionable data management plans (maDMPs)*. Research Ideas and Outcomes 3: e13086. https://doi.org/10.3897/rio.3.e13086

[63] SI was an early partner in the public launch of DMPTool.org, run by the California Digital Library.

[64] The RDMPP team created such boilerplate for staff to use in NSF grant applications, available on the Confluence site https://si-confluence.si.edu/display/RDMPP/NSF+DMP+Boilerplate

more.) A RDM program office will need to administratively support any new platforms, including assisting with account creation, running reports, and coordinating training.

There are many shared research tools that are already supported by OCIO, such as the Jira/Confluence platforms for collaboration, and other tools like Electronic Lab Notebooks used to share protocols and document methods, could also be centrally supported. Specialized services in data analysis are also currently provided by the Office of Research Computing through their data science fellows – this program should be continued until units hire their own data scientists. Additional services, such as programming or data migration, could also be either enabled by central IT staff helping units develop statements of work, or directly coordinated through OCIO using approved contractors[65] on an IDIQ.

> "All research projects are IT projects now"[66]

While this may not yet be true for Smithsonian humanities scholars, if we are to adopt a "digital first" strategy at SI, it will be true in the very near future. Assistance with IT related needs early in the research lifecycle – particularly in things like data modeling, data collection tool development, and choosing standards – has a significant impact later in the research lifecycle when analyzing, sharing, and ultimately archiving data.

Metrics

In order to provide quantitative measurements that support policies which include receiving credit for publication of datasets, SI will need to begin to track and monitor citations for data in the same way that staff publications are tracked and monitored in SRO. One key component of this effort will be the use of DOIs for datasets, and ORCID IDs for researchers. SRO staff have been using, and encouraging researchers to use, these two unique identifier systems, but this use should be reinforced by unit administration. Researchers should be encouraged to provide citations to SRO for data that is deposited in external repositories, and SRO staff should investigate ways to programmatically crawl data indices, such as DataCite, to retrieve citations for datasets connected to SI creators.

Describing data

Properly and completely describing data is critical to ensuring they are findable and re-usable. Though most researchers acknowledge providing thorough descriptions is important, resources are not always available to create metadata when it's needed. Additionally, some fields do not have established standards for data description – forcing individual researchers to devise their own, or cobble together a schema from several related standards. To enable any SI policies which support FAIR data principles, an RDM program will need to provide a combination of tools and direct services which lower the barriers to documenting and describing data.

Developing data collection tools and adopting repository platforms that prompt researchers to enter necessary metadata, e.g., by providing context-sensitive drop-down lists of terms or supplying boilerplate for certain types of commonly needed descriptions is one way to streamline metadata creation. Another way to support description is to provide direct metadata related services for research projects. This may involve having SIL or IT staff consult early on in a project to advise on standards selection where no existing standards are available. It might also involve having data managers or data curators work with researchers to add description, either in the form of a data dictionary or citation record, when a dataset or other research data outcomes are deposited. Direct services for metadata creation are resource intensive, and the level to which this service can be provided will depend on program staffing and available skillsets.

[65] Such as those on an IDIQ "indefinite delivery/indefinite quantity" contract.
[66] Scott Sillett, SCBI Migratory Bird Center, summing it all up.

Curation and Digital Preservation

Unlike analog data, e.g. paper field notebooks or hand-annotated texts, digital data cannot simply be stored with the expectation that future scholars may re-discover and use it. Digital files can be created in software that becomes obsolete over time, rendering them unreadable. Even if they are readable, they may be unintelligible to either a human or computer because the contents of the files are not well described.

Researchers will need to consult with digital curators who understand the domain-specific contents of their data and with Archives staff to help appraise their datasets before deposit; however, it may be more efficient to work with data curators earlier in the process when DMPs are being prepared. At the conclusion of a project, those DMPs can be used by curators to review the data before deposit, and set in motion the appropriate preservation processes. Someone will also need to review the data and attendant metadata to ensure they comply with SI policies and any applicable unit or field-specific standards. Because this may require domain expertise, this "quality assurance" should probably be done by data managers in consultation with data curators.

The pilot did not investigate all the service needs surrounding reformatting and curation of legacy data (both born-digital and paper) which will likely be extensive. It is hoped that such an investigation could be part of another program, specifically Digital Preservation.

Special Projects

If staffing and other resources permit, a program should also support special projects. Two potential types of projects – collaborating with external programs to enhance RDM capacity, and providing internships or fellowships to train future data curators – are detailed at the end of Appendix B.

Lost

In 2009 records from a late curator at NMNH were transferred to SIA, including research material in both paper and electronic formats. The digital data was transferred from 5.25" floppy discs when they were accessioned, as all outdated media files are, but the content is only hinted at by folder names like "Notes and Drawings." Many of the files have numbers for file extensions – SINUS.150, SINUS.260, etc. – meaningless to today's computers. Digital preservation tools couldn't determine what the original file formats were, or how to render any part of the files other than the filename.
A hint as to the contents of the files were eventually found in the paper records, which mentioned an old computer program called PC3D. This MS-DOS based software from the late 1980's was used to reconstruct 3D images from serial sections taken from instruments like CAT or NMR scanners, or scanning electron microscopes.

Because the original data was never migrated (converted to a newer version or compatible format without losing integrity) or normalized (converted to a different format that widely adopted, non-proprietary, and accessible), it can only be preserved as a "bucket of bits" in the hope that somewhere the old software exists, and digital preservationists can develop an emulator for it.

APPENDIX D: RECOMMENDATIONS FOR TECHNICAL INFRASTRUCTURE

Method

To determine the priorities for RDM technical infrastructure and next steps towards achieving strategic Institutional goals, the pilot team first reviewed existing survey and interview data from previous studies. They also talked with several researchers, labs, and departments to assess barriers to using existing sharing platforms and solicit their technical needs for data collection, analysis, sharing, and project collaboration. Using primarily interview data, the team created a functional requirements document the Institution could use to assess new repository platforms. To complete the recommendations, the pilot team did a market survey to determine what commercial off-the-shelf (COTS) and hosted platforms for data sharing are available, and performed basic testing of the top four[67] platforms, soliciting feedback from volunteer researchers and information professionals.

Overview

There are several components of a robust technical infrastructure to support RDM – adequate network throughput and pipelines for moving large files, computing power for analysis, secure storage, and file and metadata management. The lack of ability to transfer very large files across the network is a known issue for which OCIO is trying to find an affordable solution. Storage is a common concern of many research staff that were interviewed, and is an obvious concern for OCIO who are well aware of both the quantity and cost of existing storage. The full scope of SI's file and metadata management platform needs, however, are less clear.

A review of SI's technical infrastructure around research data comes as the Institution begins to implement a new five year strategic plan, a plan which includes a mandate to support sharing and archiving across a wide range of projects and to "create new digital platforms for scholars and educators to better access Smithsonian collections, research, and education resources"[68]. For researchers who are not directed by funder or project mandates to deposit in a particular external repository, SI should provide several alternatives, including a locally managed repository that provides functionality needed by the majority of projects.

Currently SI manages two local repositories that can be used for datasets – SIdora and SRO – along with the SI DAMS which manages still images, audio, and video. Each of these repositories should formally define their scope with regards to research data. SIdora is the most versatile of the SI repositories, with an ability to handle a wide variety of file formats and large datasets. However, lack of wide adoption by researchers, along with pending platform obsolescence, leads the pilot team to recommend replacing, or at least temporarily supplementing, SIdora with an equally robust but more user-friendly platform, followed by a re-evaluation in 3 years. Any repository platform SI adopts should also be evaluated for its ability to support digital preservation activities**.**

Summary

- Conduct a formal assessment of existing research data volume.
- Incorporate RDM technology needs into any technology planning, specifically large file transfer and storage.
- Supplement existing SIdora data repository with new system(s) that meet a majority of researcher needs.
- Work with SIA to explore systems and technologies that support digital preservation.

[67] Dataverse, DKAN, Figshare for Institutions, and Open Science Framework. See below for more information on each platform and how they were chosen.
[68] Smithsonian Institution. 2017

Storage

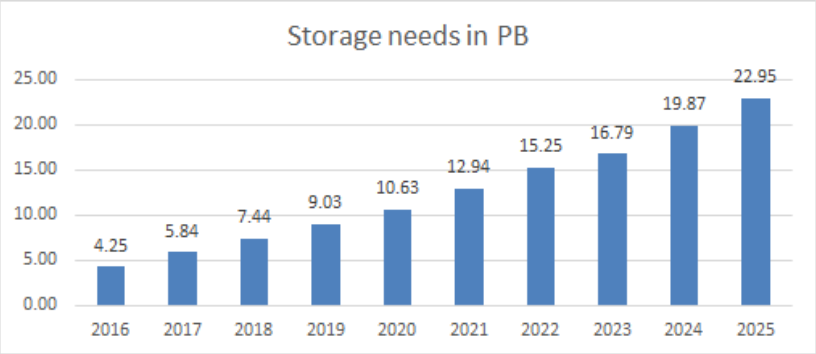
It cannot be over-emphasized that there is currently, and will continue to be, a critical need for additional secure, managed storage for research data.

> "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

Imaging data from drones and camera traps, video data from animal behavioral studies, genomics data, and oral histories are easier than ever to create, and with consumer grade electronics producing high-resolution files, the quantity of data produced by SI researchers will continue to increase rapidly. In order to plan for both current and future infrastructure needs, conducting a thorough inventory and assessment of existing research data should be a top priority for the Institution[70]. Using estimates of current and future data production from 56 survey respondents, distributed unevenly across units and research domains, the 2016 report estimated there is 6.7PB of research data at SI that needs to be managed.

> In order to quantify the need, further research to fully understand the amount of unmanaged research data and other digital assets is required.[71]

Following suggested corrections to their methodology, and eliminating the reported data from SAO (managed by SAO and NASA) the current data estimate falls to 4.25PB.[72] Reusing the reported storage growth estimates from the report as a percentage of the original deposit, gives an average of 250% growth over the first five years, and 540% growth over ten years. This results in the need for 10.63 PB by 2020 and nearly 23 PB by 2025.



Sharing and collaboration

Due to the number of high value collaborative projects, such as MarineGeo, eMammal, and Conservation Commons, and the growth of smaller multi-institution projects including those coming out of National Science Foundation (NSF) Research Coordination Network grants, there is a growing need for standard research project MOUs and well-supported collaboration tools and spaces. The pilot team believes that support for existing internally managed collaboration platforms, such as Confluence and Jira, should be formalized and expanded before being marketed to SI staff. Appropriate external platforms that support collaboration, such as Open

---

[69] Addis, Matthew. Estimating Research Data Volumes in UK HEI. figshare, 2015, 5. https://dx.doi.org/10.6084/m9.figshare.1575831.v5.

[70] This was a key recommendation from section 6.1 of the 2016 AVPreserve report

[71] AVPreserve. 2016

[72] See https://si-confluence.si.edu/display/RDMPP/Storage+Needs for calculations.

Science Framework, should also be explored. It is possible, however, that SI may be able to find or develop a platform that includes features for sharing, collaboration, and data management. Currently available platforms, however, appear to be strong in either sharing and management, or sharing and collaboration, but not all three.

Platforms for storing and sharing

Summary of functional requirements:

- Easy to get deposited data in and out, e.g., through direct download or via API
- Sufficient (but not burdensome) level of descriptive data
- Support for domain and SI metadata standards, particularly ORCIDs and DOIs
- Accepts a majority of file types produced at SI from sciences and humanities research
- Able to handle and manage large file sizes, i.e., over 2GB per file
- Ability to control access (embargo) and specify acceptable re-use (licensing)
- Compatible with future digital preservation systems

Based on comments from some researchers who were interviewed, as well as the existing technical infrastructure at SI, the team tried to include as many open source options as possible.

Preservation

Though SI does not yet have an Institution-wide digital preservation program, and the infrastructure to support digital preservation, any data platform must be able to support digital preservation activities, either through providing checksums and supporting a deposit and review process, or through connecting to dedicated digital preservation systems, e.g., Archivematica.

Description

Simply storing data, however well managed, is inadequate. In order to render the stored data findable and *usable*, it must be adequately described. Any implemented technical infrastructure must support storage of descriptive metadata along with datasets. At a minimum, a repository should support use of GUID[73]s such as DOIs and ORCID IDs, which enhance cite-ability and findability. Preferably, a system should support creation of custom fields, such as unit affiliation, to enable reporting. Any system chosen must enable the creation of metadata that meets a researcher's field, or SI's minimum, standards, but is not burdensome to use. Though metadata is vital to find and use datasets, overly complicated or time-intensive metadata forms should be avoided, lest they prove to be a disincentive for researchers to use a system.

> Continue to invest in data infrastructure that also provides rich context, detailed metadata and even a narrative account of the data creation. The kind of infrastructure researchers find most useful is where research data, papers and other outputs or resources are jointly available within a single data resource.[74]

Ideally, a platform used for RDM should be able to either virtually gather the data, descriptions, DMPs, and resulting publications in one search or be able to supply data to a portal that serves the same purpose.

[73] Globally Unique Identifiers – including Digital Object Identifiers (DOI) and ORCID ids for researchers, organizations and soon, facilities.
[74] Van den Eynden. 2014

Repositories

Currently, SI manages two local systems that can act as data repositories: SIdora, which uses the Islandora platform and SRO, which is on the DSpace platform. When looking for alternate data repositories, there are three general classes of systems to evaluate: hosted repositories[75] (both commercial and open source); locally managed repositories that are custom-built (such as the existing Islandora); and locally managed "out of the box" repositories (such as the existing DSpace.) As there are few COTS data repositories to choose from, the pilot team included several related systems that are not strictly data repositories, but can function as a "front end" to a local repository, in the initial review.[76] None of the custom solutions was tested because they require significant overhead and time to plan and build. If no other hosted or COTS systems meet all of SI's requirements, a custom solution, similar to SIdora/Islandora, should be pursued.

The pilot team reviewed the remaining COTS and hosted systems, and narrowed the list by eliminating those that could not practically be tested within the scope of the pilot (this eliminated most commercial business data systems) or that did not meet the minimum functional requirements. This left four systems, two hosted – Open Science Framework and Figshare for Institutions, and two locally installed – Dataverse and DKAN.

Dataverse[77] is an open source web application to share, preserve, cite, explore, and analyze research data. It facilitates making data available to others, and allows you to replicate others' work more easily." Originally developed at Harvard and used primarily in universities.

DKAN[78] is "a free and open source open data platform that gives organizations and individuals ultimate freedom to publish and consume structured information." It is used by many open government initiatives as a data portal, and has a science-specific module in development. It is based on CKAN[79], with a Drupal user interface.

Figshare for Institutions[80] is a version of the commonly used cloud-based commercial repository Figshare. The institutional version includes customizations to enable an organization to manage and track their researchers' outputs. There are two options for storage of data – either in the cloud or in an organization's local storage.

Open Science Framework[81] (OSF) is a free, hosted space for sharing and publishing research. Less like a standard repository than Figshare, OSF is primarily used by Social Scientists, and focuses on supporting collaborative projects. It can connect to multiple cloud storage options.

Results of repository reviews

Initial testing was done by twelve volunteers[82] including the entire MarineGEO Team at SERC (counted as one volunteer), as well as staff from SCBI, MCI, OCIO, SIA, and SIL. Staff reviewed each system and answered questions on ease of use, available features, and file handling capabilities for their specific area of research. A

[75] Hosted repositories are located and managed outside the SI network. Two hosted repositories – Elsevier's PURE and Zenodo, were eliminated from the first round of testing. PURE's terms of service are incompatible with existing SI policy. Though Zenodo is primarily Open Access, it is hosted by CERN in Europe, and adheres to European Intellectual Property laws.

[76] A complete list of reviewed repositories are on Confluence https://si-confluence.si.edu/display/RDMPP/Systems+list

[77] https://dataverse.org/

[78] https://getdkan.org/

[79] CKAN, the Comprehensive Knowledge Archive Network, is a web-based open source management system for the storage and distribution of open data and is the platform used by Data.gov.

[80] https://figshare.com/services/institutions

[81] https://osf.io/

[82] Invitations to test were sent to 40 staff from 8 units, some of whom had expressed previous interest in looking at possible repositories.

copy of the feedback form submitted by testers, along with a scored summary of test results, are available on the RDMPP Confluence site.[83] Testers were asked to provide qualitative data on each systems features, strengths, and weaknesses, with an overall system rating of 1 (bad) to 5 (great), and a yes/no question as to whether the platform should be adopted at SI.

Though the average rating of each platform was very similar, Figshare for Institutions was recommended most often as the system to be adopted at SI. Dataverse was a distant second, but had a slightly better opinion score. DKAN failed one mandatory requirement, support for multiple SI file formats. Because results were not statistically conclusive, the number of researchers surveyed was low, and some extra features were not available for testing, the pilot team recommends additional testing with researchers from across the Institution before making a final decision. Though OSF was not recommended as often, comments revealed that some testers had a very strong preference for it, indicating that OSF may fill a particular data sharing niche not covered by the other traditional repositories. As OSF is both hosted and free, the pilot team feels it should be promoted as an option for researchers, particularly those in the humanities or those involved in collaborative projects.

Dataverse, as a locally hosted repository, would take additional planning and set-up time, and require more IT resources to implement than a hosted solution like Figshare for Institutions[84]. However, because Figshare is a paid service it would initially cost more out-of-pocket but take less time and fewer staff to implement. A decision to implement either solution is of course dependent upon the availability of resources, particularly in planning, as well as funding (particularly for Figshare.) If the Institution wants to adopt any new repository system before the creation of a new RDM Program Office, staff from units outside OCIO will need to be assigned to assist with planning, implementation, and initial administration.

Discovery

Regardless of which, or how many, data repositories adopted, to improve discovery and highlight the breadth and depth of rich data resources at the Institution SI should consider creating a "data portal." This portal could include links to both collections and research data, if collecting units are willing to make their metadata openly available online[85]. Links to research data deposited in repositories outside SI will be handled through the collected data citations in SRO (see Appendix C). A portal could be built in either a tool like DKAN, or a custom portal created on top of EDAN[86], similar to the Collections Search Center.

[83] https://si-confluence.si.edu/display/RDMPP/Systems+and+IT+infrastructure

[84] For a very rough estimate of time and cost to implement, including yearly fee for Figshare for Institutions, see https://si-confluence.si.edu/display/RDMPP/System+costs+comparison

[85] As Cooper-Hewitt National Design Museum already has, through GitHub https://github.com/cooperhewitt/collection

[86] Enterprise Digital Asset Network

APPENDIX E: EXCERPTS FROM RECOMMENDATIONS OF PREVIOUS STUDIES AND REPORTS

Stewarding the Invisible: Setting the Stage for Institution-Wide Digital Preservation at the Smithsonian. AVPreserve. November, 2016

This report to the Smithsonian Digital Preservation Working Group was the first Institution-wide assessment of digital life-cycle management practices at SI and a readiness assessment for the Institution's ability to implement systematic digital preservation activities. The extracted recommendations below, though directed more broadly to all digital assets at SI, apply particularly to digital research data.

> DEFINE ROLES AND RESPONSIBILITIES (p.77)
>
> Roles and responsibilities should eventually be formalized in Smithsonian Directives, and should include….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 [or IT] staff they can turn to for help.
>
> FORMALIZE TERMINOLOGY (p.79)
>
> For research data, define the categories of data…. 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.
>
> CREATE A SMITHSONIAN DIRECTIVE FOR RESEARCH DATA MANAGEMENT (p80)
>
> …. 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.
>
> CONDUCT TRAINING AND OUTREACH (p.82)
>
> 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….Outreach would include workshops, tutorials, and consulting for existing staff, as well as new staff and research fellows as part of their onboarding process.
>
> Establish, track, and report on metrics that illustrate (the value of digital preservation) (p.83)
>
> Expand [the Smithsonian Institution Dashboard] to include metrics that …. 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 research data available to the public; amount of data repurposed for new research; number of citations of publicly accessible datasets; or number of peer-reviewed publications based on data generated by Smithsonian researchers.
>
> CLARIFY THE ROLE OF EXISTING REPOSITORIES (p.83)

> We recommend that each existing repository establish its scope, and where gaps exist, the repository managers, in coordination with the [a program office], must decide whether to update existing policies to include orphan resource types, or create new repositories that will take responsibility for them….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.
>
> GATHER REQUIREMENTS FOR RESEARCH DATA INFRASTRUCTURE (p.84)
>
> …the input gathered during our interviews indicates that this repository [SIdora] may not be meeting current needs. We recommend that rather than continuing to build out this service, pause development and initiate a comprehensive requirements development process. The outcome of this effort may indicate that SIdora needs to be reimagined, re-branded, or rearchitected. There is the potential, as well, the research finds that a new system should replace it.
>
> MOVE AWAY FROM RELIANCE ON PROJECT-BASED FUNDING (p.86)
>
> Sustainability and persistence requires …. 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 unfindable), research data. …. 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.

Sharing Smithsonian Digital Scientific Research Data from Biology. SI Office of Policy and Analysis. March 2011

At the request of the Office of the Under Secretary for Science, this study examined how and to what extent the Smithsonian shares biological research data both within and outside the Institution.

> FROM THE INTRODUCTION (p.1)
>
> Easy access to a wide range of usable biology data remains an elusive ideal. Achieving this ideal will require expensive technological infrastructure; new administrative policies, structures, and workflows; specialized teams of personnel that combine domain research, information-technology (IT), and information-management skills; and, ultimately, coordination and collaboration across organizations and governments.
>
> It will also require changes in the biology research culture…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. Professional disincentives to (and a parallel lack of incentives for) data management and sharing reinforce this norm; these include the emphasis on publishing peer-reviewed articles and the absence of professional credit for producing, curating, and sharing valuable data sets per se. One result has been the creation of a great many disconnected data sets at a great many locations in myriad forms. These can be very difficult to discover, access, and use, and the already-enormous volume of such data is growing exponentially. Some are already past the point of use and preservation, and many more are at risk of permanent loss.
>
> FROM THE CONCLUSIONS: (p.109-121)

CONCLUSION 1: The small-science[87] approach to the management and sharing of digital biology research data is anachronistic. It is at variance with the growing emphasis both in U.S. policy and around the world on open access to scientific data, and it may put the results of important research investments at risk and impede long-term access to valuable scientific resources.

…the Smithsonian needs to carry out systematic and thorough management of its biology data throughout their lifecycle. While the study team encountered…important initiatives at the Smithsonian to further systematic data management and sharing, it also found that an Institutional strategy to guide progress in these areas is lacking. As a result, discovery, access, and use of Smithsonian biology data are constrained, and an ever-increasing volume of legacy data is at risk. Going forward….the loose data-management norms associated with small science will have to give way to more standardized, systematic methods.

CONCLUSION 2: To make its digital biology data easily discoverable, accessible, and usable by internal and external users, the Smithsonian needs to unequivocally articulate a policy of open access and systematically establish the capacity and tools to implement that policy. The study team believes that the Smithsonian, as a largely taxpayer-funded entity, has an obligation to provide open access to its biology data, subject to reasonable proprietary waiting periods and other justifiable exceptions….[make] open access a fundamental operating principle of the Smithsonian's research enterprise and establishing external usability as a primary consideration in decisions regarding data-management processes, standards, infrastructure, and technology.

CONCLUSION 3: Sharing of Smithsonian biology data requires fundamental changes in current data-management and -dissemination practices. The current small-science, seat-of-the-pants, fragmented, and mostly unit- or department-based approach will have to give way to a set of core Institution-wide principles and standards. ….The critical elements that need to be included in Smithsonian-wide policy include: data management and sharing standards, compliance with those standards, a central record of data holdings, a tool to facilitate discovery and access, a trusted digital repository.

CONCLUSION 4: Systematically and immediately addressing the risk of legacy data loss and preventing further growth in the backlog of Smithsonian legacy data are high priorities. Two steps to minimize the further growth of legacy data are to 1) ensure that the data of researchers soon to retire, and projects soon to end (or recently ended), receive near-term basic data management and are transferred to secure storage; and 2) require that scientists with ongoing projects list their data in the central record of Smithsonian holdings and routinely back them up on a stable medium.

CONCLUSION 5: It is important that the central administration be proactive in reaching out to Smithsonian biology researchers … and obtain buy-in among research staff for enhanced data management. Researchers currently have virtually no incentives, and many disincentives, for engaging in data management beyond the minimum required …Of particular importance to such an effort are: Inclusion of professional credit for effective data management in performance evaluations…Providing researchers with access to support personnel and tools that facilitate data management and minimize the time researchers need to devote to it; and access to funds earmarked for data management and additional data-management support staff.

[87] "In this report, small science refers to fields in which digital data-management practices tend to be driven by the needs of individual, often small-scale, research projects, rather than by standards common to a whole field. Big science fields, by contrast, have standards of data management that are widely accepted and used by researchers." P. 2

CONCLUSION 6: Meeting the growing challenges of digital data management and sharing at the Smithsonian will require additional resources. Additional data-management support staff are an especially critical resource need. Effective data management and sharing require a cadre of research support personnel who combine IT, information-science, and domain-science expertise. The allocation of such personnel between research and central support units remains an open question.

CORE RECOMMENDATIONS (summarized)

1. Commit to open access for data, with external usability a primary consideration regarding data management processes, standards, infrastructure, and technology.
2. Establish capacity and tools to make data discoverable, accessible and usable .
3. Engage with external organizations working to advance data sharing and management.

Harnessing the Power of Digital Data for Science and Society. National Science and Technology Council – Interagency Working Group on Digital Data. January 2009

RECOMMENDATIONS (EDITED)

…Appropriate departments and agencies [should] lay the foundation for agency digital scientific data policy and make the policy publicly available.

….The goals of the agency data policy should be to maximize appropriate information access and utility and to provide for rational, cost efficient data life cycle management. Agency data policies should be publicly available and should guide and inform the development and implementation of data management plans in individual projects and activities.

(3) …Agencies promote a data management planning process for projects that generate preservation data.

In particular, agencies could consider requiring data management plans for projects that will generate preservation data …. Data management plans should provide for the full digital data life cycle and should describe, as applicable, the types of digital data to be produced; the standards to be used; provisions and conditions for access; requirements for protection of appropriate privacy, confidentiality, security, or intellectual property rights; and provisions for long-term preservation (including means for continuously assessing what to keep and for how long).

LIST OF ABBREVIATIONS

COTS – Commercial off the shelf

DMP – Data Management Plan

DOI – Digital Object Identifier

FAIR – Findable, Accessible, Interoperable, and Reusable

GUID – Globally unique identifier

IDIQ – Indefinite delivery/indefinite quantity

MCI – Museum Conservation Institute

MOU – Memorandum of Understanding

OCIO – Office of the Chief Information Officer

OPandA (or OP&A) – Office of Policy and Analysis, former name of SOAR

RDM – Research Data Management

SCBI - Smithsonian Conservation Biology Institute

SERC - Smithsonian Environmental Research Center

SIA – Smithsonian Institution Archives

SIdora – The SI data repository

SIL – Smithsonian Libraries

SOAR – Smithsonian Organization and Audience Research

SOW – Statement of work

SRO – Smithsonian Research Online

SELECT BIBLIOGRAPHY

AVPreserve. *Stewarding the Invisible: Setting the Stage for Institution-wide Digital Preservation at the Smithsonian.* 2016

Bryant, Rebecca, Brian Lavoie and Constance Malpas. 2017. *Scoping the University RDM Service Bundle. The Realities of Research Data Management, Part 2*. Dublin, OH: OCLC Research. doi:10.25333/C3Z039

Fearon, David Jr., Betsy Gunia, Sherry Lake, Barbara E. Pralle, and Andrew L. Sallans. *Research Data Management Services.* SPEC Kit 334. Washington, DC: Association of Research Libraries, July 2013. https://doi.org/10.29242/spec.334

FORCE11. Guiding Principles for Findable, Accessible, Interoperable and Re-usable Data Publishing. 2014. https://www.force11.org/fairprinciples accessed 2/20/2018

Jones, S., Pryor, G. & Whyte, A. *How to Develop Research Data Management Services - a guide for HEIs*. DCC How-to Guides. 2013. http://www.dcc.ac.uk/resources/how-guides/how-develop-rdm-services

McGinty, Stephen, Rebecca Reznik-Zellen and Jessica Adamick. 2012. *Tiers of Research Data Support Services*. Journal of eScience Librarianship. http://dx.doi.org/10.7191/jeslib.2012.1002

National Academy of Sciences (US), National Academy of Engineering (US) and Institute of Medicine (US) Committee on Ensuring the Utility and Integrity of Research Data in a Digital Age. *Ensuring the Integrity, Accessibility, and Stewardship of Research Data in the Digital Age*. Washington (DC): National Academies Press (US); 2009 https://www.ncbi.nlm.nih.gov/books/NBK215270/

OECD. 2017. *Business models for sustainable research data repositories*, OECD Science, Technology and Industry Policy Papers, No. 47, OECD Publishing, Paris, http://dx.doi.org/10.1787/302b12bb-en .

Smithsonian Institution. Office of Policy and Analysis. *Sharing Smithsonian Digital Scientific Research Data from Biology*. March 2011. https://repository.si.edu/bitstream/handle/10088/26322/11.03

Smithsonian Institution. *One Smithsonian: Greater Reach, Greater Relevance, Profound Impact.* Smithsonian Institution Strategic Plan 2017-2022.

Smithsonian Institution Libraries. Data Management Working Group. May, 2014. *Recommendations for Libraries' Involvement in Data Management at SI*. Internal memo, available on request.

Van den Eynden, V. and Bishop, L. *Incentives and motivations for sharing research data, a researcher's perspective.* A Knowledge Exchange Report. 2014. http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf . 2014. accessed 11/2017

Vines, Timothy H. et al. *The Availability of Research Data Declines Rapidly with Article Age*. Current Biology , Volume 24 , Issue 1 , 2013 https://doi.org/10.1016/j.cub.2013.11.014

Wilkinson, M. D. et al. *The FAIR Guiding Principles for scientific data management and stewardship*. Sci. Data 3:160018 2016. doi: 10.1038/sdata.2016.18

CREDITS

The Research Data Management Program Pilot team was:

Keri Thompson, SIL
Beth Stern, OCIO
Lynda Schmitz Fuhrig, SIA
Sue Zwicker, SIL

The pilot team would like to thank the Steering Committee who supported our efforts

Scott Miller, DUSCIS; Martin Kalfatovic, SIL; Deron Burba, OCIO; Nancy Gwinn, SIL; Anne Van Camp, SIA

And our supervisors, who allowed us to take time from our regular duties to work on this important project.

We would also like to thank all the reviewers and staff who contributed to the final report, particularly:

- Gale Robertson, NMNH Education, who created the policy review matrix as part of her ELDP rotation.
- Joanna McCloud, Data Curation contractor, who managed system testing and review.

Biggest thanks of all to the researchers and other staff at SI who generously shared their time and expertise with us, especially all our system testers.