Smithsonian Data Management Best Practices

Planning and data management plans

The following best practices are intended for use by Smithsonian researchers and affiliated staff who plan for, create, and/or work with digital research data. Additional information about available tools, policies, and resources for managing research data can be found on https://library.si.edu/research/data-management.

In this document

Image: Operation Reindeer SIA2011-0030
Courtesy Smithsonian Archives

Many granting bodies, such as NSF and the Alfred P. Sloan Foundation, require a formal **data management plan (DMP)**[i] as part of a grant proposal.

Even if a proposal does not require a DMP, SI strongly recommends that PIs create a planning document before starting any project that will create digital research data[ii]. DMPs are valuable tools for addressing issues that affect not only collection and use, but also the long-term viability of your data.

A written data management plan can:

- provide **continuity** on projects if staff join or leave
- allow for future **validation** or reproduction of results
- enable **reuse** of your data in potentially novel ways

SI Libraries staff are available for consultation on creating DMPs and are happy to review draft DMPs before submission with a proposal. Contact AskaLibrarian@si.edu for more information.

PROPOSALS

The Smithsonian Office of Sponsored Projects (OSP) provides administrative and financial services for externally funded grants and contracts, and is available to assist PIs with technical and procedural questions related to managing grants and awards.

OSP also provides training in proposal development, writing and editing, and compliance oversight for areas such as Institutional Animal Care & Use, Export Control, Human Subjects in Research, and Responsible Conduct of Research. Their list of online and in person learning opportunities are available on their PRISM site.

PLANNING CHECKLIST

Any plan should *at a minimum* answer the following questions in **bold** for each stage in the data management lifecycle. More specific guidance for questions in the data collection, publishing, and archiving stages is available at https://library.si.edu/research/data-management

Proposal/planning Stage

- ☐ **What type of data is being collected/generated?**
- ☐ **Who is involved in data collection?**
- ☐ **Who "owns" the rights to the data?**
- ☐ **Are there restrictions on sharing and reuse?**
- ☐ **Are there applicable institutional policies on how the data is handled, shared, or archived?**
- ☐ Who will be using the data?
- ☐ If a collaborative project, are there MOUs that define roles and responsibilities?
- ☐ How do the outcomes need to be reported, e.g., to a sponsor or publisher?

Data Collection Stage

- ☐ **How will data be acquired/collected?**
- ☐ **What metadata[iii] standards and schema[iv] will be used?**
- ☐ **What are the file and data field naming conventions?**
- ☐ **What are the temporary storage requirements (size, cost, media)?**
- ☐ **How, where, and how frequently will data be backed up?**
- ☐ Are there existing standards for data structure and vocabularies, or will they be developed?
- ☐ Are there existing workflows or tools for collecting, processing, describing, and storing the data, or will they need to be developed?
- ☐ Is there a data model[v] for the project?
- ☐ Will your data be versioned[vi], and if so, how will versioning be handled?
- ☐ What is your quality assurance/quality control[vii] process?

Publishing Stage

- ☐ **What repository[viii] or platform will be used to share the data?**
- ☐ **Who will be responsible for deposit and archiving[ix] after the project ends?**
- ☐ **If the data is to be shared publicly, what license should be applied? Are there any use restrictions?**
- ☐ If the data is embargoed, what is the embargo period, and who will manage it?
- ☐ If the data is not public, how will access be restricted?
- ☐ What costs are associated with publishing?
- ☐ What unique persistent identifier[x] will be assigned to the data (DOI, etc.)?

Archiving Stage

- ☐ **Who is responsible for preserving the datasets in the future?**
- ☐ **What data should be retained?**
- ☐ **Where will the data be archived?**
- ☐ **How much storage will be needed?**

Last update: 2018-03-12

- ☐ **How long should the data be maintained, and why?**
- ☐ **What are the risks for future access to the data, i.e., proprietary file formats, specialty software needed to interpret, password-protected systems?**
- ☐ How should the data be maintained in the future?
- ☐ Is there a cost associated with archiving the data?
- ☐ How will the data be found?

Funder-specific DMP Requirements

Some funding agencies require that plans submitted with grant proposals include specific elements or specific formatting. Below is a list of links to those requirements, alphabetical by funder, for selected granting organizations.
* = sample plans available on their site

- Alfred P. Sloan Foundation
- BCO-DMO NSF OCE: Biological and Chemical Oceanography
- Department of Energy – DOE: Generic
- *Gordon and Betty Moore Foundation (pdf)
- Institute for Museum and Library Services IMLS : guidelines for datasets (Word doc)
- *National Aeronautics and Space Administration NASA
- *National Endowment for the Humanities NEH-ODH (pdf)
- National Oceanic and Atmospheric Administration NOAA
- *National Science Foundation NSF-Generic
  - NSF-Atmospheric and Geo
  - NSF-Astronomy (pdf)
  - NSF-Biology (pdf)
  - NSF-Earth Sciences
  - NSF-Education and Human Resources (pdf)
- United States Geological Survey USGS

TOOLS AND TEMPLATES

Smithsonian Specific Templates

The Data Management Team has developed boilerplate (temporarily located on an internal Confluence site) that can be used when applying for an **NSF** grant if you plan to deposit data either in **SRO**[xi] or **SIDora**[xii]**.** The boilerplate address the specifics of data storage and archiving, dissemination, policies, and roles and responsibilities within the SI data management ecosystem.

DMPTool

One of the major tools for creating data management plans is the DMPTool, hosted by the University of California Curation Center (UC3). The Smithsonian was one of the original partner institutions involved in creating the DMPTool.

The DMPTool website includes templates and requirements for a large number of granting bodies, including NSF, DOE and NIH.

> Any researcher at the Smithsonian can create an account and login to the DMPTool by selecting "Smithsonian Institution" from the list of institutions and then using their **SI network** username and password.

Digital Curation Centre Checklist

The Data Curation Centre's (DCC) Checklist can help you craft a custom DMP. The Checklist covers the main elements of a good plan, most of which are listed above, with suggested content for each element.

Last update: 2018-03-12

[i] A data management plan is a formal document that outlines what you will do with your data during and after a research project. DMPs help articulate who has responsibility for your data throughout its lifecycle, what can and cannot be done with the data, as well as how to find, re-use, and interpret your data. Many potential data management disasters can be handled easily or avoided entirely by planning ahead.

[ii] The primary source materials that support research, scholarly enquiry, or artistic or technical innovation. Research data are used as evidence in in the research process and/or are necessary to validate research findings and results.

[iii] Metadata is descriptive "data about data." This term can be used in many ways—in the data dictionary, metadata provides descriptions of and context for the contents of your dataset. In a dataset citation record, the metadata provides a high-level description of the overall dataset.

[iv] Metadata schema are a standard framework or set of rules that are used to provide a consistent, structured description of an object, e.g., DarwinCore, schema.org.

[v] A data model defines how the different elements (measurements, data points, files, etc.) of your data are connected, and how they interact. Creating a data model can be very helpful when planning how to collect and later display large amounts of information.

[vi] Versioning is the process of saving, and then assigning unique names or numbers, to a file whenever significant changes are made. Versioning files enables tracking changes over time, as well as supporting reverting to an older version of a file. Several software tools and file systems support automatic versioning including: DropBox, Git, ArcGIS, and most wikis.

[vii] Though often used interchangeably, quality assurance (QA) and quality control (QC) are activities invoked at different stages of a project. QA establishes processes and standards to ensure data is collected or created in a way that meets project requirements, i.e., error prevention. QA may involve using data collection forms with drop-down menus instead of free-text fields, or writing protocols that specify use of particular measuring tools. QC establishes processes to check for compliance with project requirements, and specifies how to correct or normalize data that does not meet those requirements, i.e., error detection and correction.

[viii] A repository is an actively managed system used to store scholarly output for long-term preservation. A repository may or may not be publicly accessible. A web server, external hard drive, or cloud storage system such as Google Drive, are not repositories.

[ix] Archiving here refers to depositing data in a managed system and service that organizes and preserves information for future retrieval and use.

[x] A Persistent identifier is a stable alphanumeric or numeric descriptor attached to a digital object that is used to uniquely identify it. A GUID, or Globally Unique ID, is a type of persistent identifier that is unique within a large identifier ecosystem. Common types of persistent identifiers used in data management include DOIs, handles, ORCiDs (for authors), ARKs (Archival Resource Keys), and EZIDs.

[xi] Smithsonian Research Online (SRO) is managed by Smithsonian Libraries and has two primary components: a bibliography of publication citations and a repository of full-text online editions built on the DSpace platform. While the repository is primarily used for publications, it can also accommodate datasets and other supplementary material. More information can be found at https://research.si.edu/about/ .

[xii] SIDora is the Smithsonian data repository. Managed out of the Office of Research Computing, OCIO, SIDora is built on a Fedora Commons 3 platform, with a Drupal front end. It can accommodate data projects that are large, have custom interface needs, or are structurally complex. More information can be found at https://oris.si.edu/sidora