Smithsonian Data Management Best Practices

Choosing a Repository[i]

Given the large number of specialty repositories that exist or are being built for specific data types, specific organisms, and large grant-funded collaborative projects, it is impractical to list all the data repositories that could be used by SI researchers to conform to the FAIR[ii] (Findable, Accessible, Interoperable and Reusable) data principles.

Before depositing data in a repository not included in one of the lists below, the data owner *should at a minimum* ensure that the repository:

- Has a plan and sufficient funding to ensure its **long-term viability**.
- Allows **export** of data and data descriptions in a standards-compliant format, preferably identical to the format you deposited.

Ideally, the repository should also:

- Enable easy citation of your data, including supporting DOIs[iii] (either minted by the repository or by SIL).
- Be searchable, and indexed in a service such as DataCite or Elsevier's DataSearch.
- Support application of an appropriate license, and embargo of data if necessary.
- Support metadata standards for your data, e.g., ISO 19115 for Geographic data.

RE3DATA.org - a registry of research data repositories - is an excellent source of detailed information on individual repositories.

In general, any repository managed by a U.S. **Federal agency or national laboratory**, e.g., NIH's GenBank, NASA's National Space Science Data Center, or ORNL's DAAC should be considered a **preferred repository** for any SI research data that meets their criteria for deposit. Data repositories run by established U.S. institutions such as Harvard's Dataverse are also acceptable.

General–purpose

SI has two repositories, SRO and SIdora, that accept Smithsonian-produced data.

Both SRO and SIdora support multiple file types, and are discipline agnostic. Both accommodate use of DOIs for citation. Both have actively managed, backed-up, secure storage in the Herndon Data Center. Both support having open (accessible) and closed (private) data, though SRO additionally supports embargoes and other restrictions.

- **SRO**[iv] – is best for smaller (<50GB), fixed (inactive) datasets that accompany or support publications deposited in SRO.
  →To deposit data and publications in SRO, you can self-deposit using the forms found here http://staff.research.si.edu/input_forms.cfm or contact research-online@si.edu

- **SIdora**[v] – is best for larger, or more complicated datasets, including actively updated datasets
  →To deposit data in SIdora contact Beth Stern or email si-sidora@si.edu

If you or your publisher prefer to deposit with a non-SI repository, there are four recommended general-purpose repositories. Their features are compared below.

feature	**Dryad**	**Figshare**	**Open Science Framework**	**Zenodo**

	http://datadryad.org/pages/faq	https://support.figshare.com/	http://help.osf.io/	http://help.zenodo.org/
Fees/terms	$120/deposit	free, premium service for a fee	free	free, limited to 50GB/dataset
dataset licensing	CC-0 (Creative Commons 0 only)	CC-By, CC-0, MIT, GNU GPLv3,Apache 2.0	CC, Apache, MIT, GNU, other	CC, other
access options	open; embargoed (only for certain publishers)	open; restricted (unpublished)	open; restricted; closed	open; restricted; closed; embargoed
versioning[vi] available	yes	yes	yes	yes
formats accepted	office documents, scientific & statistical data, plain text, structured text, software, source code, other	office documents, images, structured graphics, audiovisual data, raw data, plain text, archived data	any	any
provides usage stats	yes	yes	yes	yes
persistent identifiers[vii]	will assign DOI	supports ORCID; will assign DOI	supports ORCID; will assign ARK and DOI	supports ORCID; will assign DOI or use provided DOI

Discipline-specific

The following non-comprehensive list does *not include* repositories for model organisms, those managed by U.S. government agencies, or institutional data repositories such as Harvard's Dataverse or UIUC's Illinois Data Bank. *Omission from this list does not imply that it is not an acceptable repository for your data*. If you would like to see a repository listed here, contact Keri Thompson thompsonk@si.edu.

- Astronomy
  - SIMBAD http://simbad.u-strasbg.fr/simbad/
- Biodiversity/taxonomy
  - GBIF Global Biodiversity Information Facility https://www.gbif.org/
  - KNB Knowledge Network for Biocomplexity https://knb.ecoinformatics.org/
- Omics/sequencing
  - ArrayExpress https://www.ebi.ac.uk/arrayexpress/
  - BioGRID https://thebiogrid.org/
  - DGVA Database of Genomic Variants Archive https://www.ebi.ac.uk/dgva
  - NURSA Nuclear Receptor Signaling Atlas https://www.nursa.org/nursa/index.jsf
  - ProteomeXchange http://www.proteomexchange.org/
  - UniProt https://www.ebi.ac.uk/uniprot
- Environment/Climate
  - NERC Environmental Information Data Centre (multiple repositories) http://www.nerc.ac.uk/research/sites/data/
  - Environmental Data Initiative https://portal.edirepository.org/nis/home.jsp
  - World Data Center for Climate (at DRKZ) https://www.dkrz.de/up/systems/wdcc

    - PANGAEA https://www.pangaea.de/
- Humanities
    - ICPSR https://www.icpsr.umich.edu/icpsrweb/landing.jsp
    - Qualitative Data Repository (Political Science) https://qdr.syr.edu/
- Marine Sciences
    - Seanoe http://www.seanoe.org/
- Science (General)
    - Mendeley https://data.mendeley.com/faq

References

Fairsharing.org : PLOS recommended repositories. https://fairsharing.org/recommendation/PLOS . Accessed 2018-02-01

Re3Data : Registry of Research Data Repositories. http://re3data.org Accessed 2017-11-02

[i] Repository - an actively managed system used to store scholarly output for 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.

[ii] FAIR data principles – FAIR is an acronym for Findable, Accessible, Interoperable, and Re-usable. The principles are meant to guide data producers and publishers as they strive to practice good data management and enable leveraging high quality formal digital publications to facilitate future reuse, discovery, and evaluation.
Read all the principles on the FORCE11 website: https://www.force11.org/group/fairgroup/fairprinciples

[iii] DOI – A digital object identifier (DOI) is a very specific kind of persistent identifier. DOIs are assigned to a digital object like a publication or dataset to facilitate finding and citing that object. DOIs are "resolvable" – that is, if the object moves to a new online location, the registry that tracks that object can be updated to make sure any links to that object do not break.
Many repositories will issue DOIs for deposited data. The Libraries can also issue a DOI for a dataset or publication that is created by SI staff, and stored in an SI managed location such as in SRO, on an si.edu website. All digital objects with DOIs issued by the Libraries are included in metrics and impact reports submitted to the Castle.

[iv] 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/ or contact Alvin Hutchinson hutchinsona@si.edu

[v] SIDora – 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 or contact Beth Stern sternb@si.edu

[vi] Versioning – 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] Persistent identifier (PID) - 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.