Caltech Library

Data Management Plan Example

This example is based on the generic NSF data management plan template. Always refer to the specific guidance for your grant proposal when writing a data management plan, as requirements differ.

Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project

This project will produce 2-, 3-, and 4-dimensional microscopy images of mouse cells in the TIFF image format. Individual images will be anywhere for a couple GB to hundreds of GB in size. We expect to produce 50 TB of image data over the course of the entire project, and this proposal includes funding to store all data for three years past the tenure of the grant.

The project will also create software code in the Python programming language for the analysis of the microscopy image files.

Standards to be used for data and metadata format and content (where existing standards are absent or deemed inadequate, this should be documented along with any proposed solutions or remedies)

Images will be saved in the native file type of the microscope, Zeiss LSM, and converted to the OME-TIFF format (https://docs.openmicroscopy.org/ome-model/6.0.1/ome-tiff/specification.html) using the Bio-Formats tool. OME-TIFF images use the OMEXML metadata standard to document images. We will additionally document our data in lab notebooks and digital README.txt files.

All files will be named using the form sampleID_YYYYMMDD_R#.ome, where R is the revision number. Each sample will be assigned a unique sampleID that will connect to information about the sample in the README.txt files. Once files are ready for publication, project members will create DataCite standard metadata that describes the files generally in the CaltechDATA repository.

Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements

All raw data will be stored locally and backed up on a central server while data is being collected and analyzed. The PI holds final responsibility, but is active in emphasizing the need for regular data backup and storage within the group. Group members are taught the naming scheme and file organizational system for the laboratory and are expected to maintain these. Through remote storage, a shared drive, and shared desktops in the laboratory, students create regular backups of their files, which are checked by the PI.

Caltech Library

Data Management Plan Example

Once results are ready for publication, data files associated with the results will be uploaded to the Caltech Research Data Repository (data.caltech.edu). CaltechDATA is a free service for all members of the Caltech community for storage of up to 500 GB of data, which will be sufficient for the public data generated from this project. Each data set will be assigned a digital object identifier (DOI). All data will be preserved at least 5 years past the tenure of this grant. All publications resulting from this grant will include the data set DOI(s) that support that work. In the unlikely event that CaltechDATA becomes unavailable, an alternative repository such as Zenodo or the Harvard Dataverse will be employed.

All software generated by this project will be managed with the version control software git and shared publicly on GitHub. All software will follow a standard release cycle, and each release will be assigned a DOI and preserved in the CaltechDATA repository.

Policies and provisions for re-use, re-distribution, and the production of derivatives

All publicly shared data will be licensed with a creative commons zero (CC-0) license. This dedicates all data to the public domain and allows for unrestricted re-use. It is expected that users that employ a data set generated from this grant will provide a citation.

Caltech holds the copyright in software generated under sponsored research projects, but software from this project will be made publicly available under a standard BSD-3-clause license. Use of standard licenses means that other users and machines have a clear understanding of the conditions of re-use for all data and software.

Plans for archiving data, samples, and other research products, and for preservation of access to them

All publicly accessible data described above will be archived on CaltechDATA, which is managed by the Caltech Library. All data is preserved using the OAIS reference model and multiple copies of all data sets are stored at the Caltech Library and remotely. CaltechDATA is intended as a long-term data repository and it is expected that data will be available for the foreseeable future. At a minimum, we guarantee all public data will be available for 5 years past the end of this grant. Public data can also be distributed automatically to interested disciplinary repositories via the CaltechDATA API or DataCite API.

All data files uploaded are human-readable and suitable general format that does not depend on a specific analysis program to be usable. In addition, all data will be retained locally with multiple backup hard drives within the lab.