# EXAMPLE DATA MANAGEMENT PLAN: BIOLOGY

**This data management plan (DMP) is part of a collaborative proposal that is expected to run for three years.**

## **Research products and data formats**

Experimental data collected as part of this project will consist of time-series recordings of images or voltages. Computational data will be water flow velocity and pressure on two- or three-dimensional grids, along with positions of the animal and stimulus. As much as possible, we aim to store data in the Hierarchical Data Format, version 5 (HDF5; for specifications, see <http://www.hdfgroup.org/HDF5/>), a general purpose open-source scientific data format that allows both data storage and annotation. Standard annotation will include sampling rate, scales and calibrations, raw data units (e.g., volts), and scaled data units (e.g., stimulus vibration frequency).

**Primary experimental data** The following primary data are anticipated. Unless otherwise noted, data are stored or converted into HDF5:

1. Voltage data including (a) recordings of lateral line nerve activity (§5.1.1), used to examine how the lateral line responds to self-noise and external stimuli; (b) simultaneous recordings of stimulation in the mesencephalic locomotor region (§5.1.3), used to generate naturalistic swimming; (c) recordings of the motor position used to produce swimming-like motions (§5.1.4); (d) driving voltages for the external stimulus (§5.1.5). Data are initially acquired and stored using LabChart Pro (<https://www.adinstruments.com/products/software/labchart-pro>) and then converted to HDF5 using a custom converter.
2. High speed video recordings of the motion of the animal’s body for determining the kinematics of the motion (§5.1.4). Video is sampled at 100-200Hz and stored as uncompressed AVI files or as HDF5 files.
3. Laboratory notebooks and other notes. These are stored electronically using the LabArchives software (http://www.labarchives.com/), hosted on the Tufts UIT Research Technology Services’ central archive (uit.tufts.edu/at/?pid=80). PDF backups will be made periodically.

**Primary computational data** Computer simulations will produce estimates of two or three-dimensional flow and pressure fields (§5.2), along with positions of the fish and/or stimulus, and the state of the neuromast (§5.2.3) and hair cells and afferent nerves (§5.2.4). Data generated through computer simulations is expected to complete ~60 fully validated production runs which will produce archival quality data. Each of these simulations, which runs over time-periods extending from days to weeks, will produce about 50 GB of uncompressed data.

**Secondary/processed data** Additionally, these raw data will be processed using several software packages, primarily Matlab (MathWorks) that is used for most secondary analyses. The latest version of Matlab data files are compatible with HDF5, and so no conversion is necessary. Matlab provides poor abilities to annotate data files; instead we will aim to develop documented analysis functions that will perform the analyses on the primary data and produce the secondary data files along with figures where appropriate.

**Computer code and algorithms** The software developed in this project includes the various simulation code and post processing codes. Software will be primarily developed in Fortran 90 and MATLAB.

**Metadata** Metadata will be provided.  This meta-data will be obtained by several means: (a) The software will automatically generate, along with the simulation results, the metadata for the outputs. The metadata includes the file name for the results, location, owner, create time, model version, runtime, inputs, hardware information, etc. The output filename will appear as ModelName-VerNo-Date(YYYYMMDD)-Time(HHMMSS); (b) The project will document information about the context, content, quality, provenance, and/or accessibility of the data used. This will also include information embedded in HDF5. Additionally, the project will seek to document information about authors, dates and brief descriptions for scanned PDFs, notebooks and lab work.

## Data storage and archiving

**Storage and Backup** Raw and processed data and computer code will be archived by the PIS on Tufts University and Johns Hopkins University servers for a period of not less than five years after completion of the project. The electronic lab notebook (LabArchives) will also be stored on the Tufts servers as an index and description of the data files. Tufts provides centralized, backed-up, network accessible storage for research projects without cost. Tufts University maintains a distributed information technology environment, with central as well as local aspects of overall planning and control. Tufts' information security program is structured in a similar manner. Operationally, Tufts central IT organization (UIT) and each local IT group maintain standards of quality and professionalism regarding operational processes and procedures that enable effective operational security. The project data will be archived in collaboration with Tufts University Information Technology (UIT). Research data for this study will be stored on Tufts University’s centrally provided “shared research storage facility”, which is provisioned on TuftsNetApp model 6040 enterprise networked storage. Provisioning occurs through self-service at <http://researchstorage.uit.tufts.edu>. All devices and users are subject to the Tufts Acceptable Use policy, which can be accessed at <http://uit.tufts.edu/?pid=786>. The system administrators at JHU who manage the clusters where these simulations are run (see Facilities and Resources Document) in this project will provide data backups. The backups are expected to take place daily/weekly/monthly and will be stored locally at JHU.Access to data and data sharing policies

**Data Capacity and Volume** Many dozens of simulations will be initiated each year as part of this project, but many of these simulations will either not be successful (due to accuracy, convergence or stability issues) or will be used to gain understanding for enabling simulations with acceptable quality. In the end, we expect that we will complete 20 simulations each year which have adequate accuracy, convergence and quality to have archival value. Each of these simulations generates about 50 GB of data.

**Long Term Archiving and Preservation**  Two years after publication or after completion of the project, raw data and analysis code will be made available to the community. The PIs will retain principal legal rights to the intellectual property developed under this grant as is compliant with the “Tufts University’s Policy on Rights and Responsibilities with Respect to Intellectual Property” ([viceprovost.tufts.edu/research-policies/intellectual-property](file:///C%3A%5CUsers%5Crraboi01%5CAppData%5CLocal%5CMicrosoft%5CWindows%5CTemporary%20Internet%20Files%5CContent.Outlook%5C2CVSS2BA%5Cviceprovost.tufts.edu%5Cresearch-policies%5Cintellectual-property)). Selected code will be distributed publicly on the PIs’ websites at Tufts and JHU. Web pages containing an indices of data will be developed and posted publicly as a link from the PI’s website. Data will be shared on an *ad hoc* basis to researchers who contact the PIs. Additionally, data will be deposited as appropriate in a centralized repository, to be determined (such as http://datadryad.org/).

**Data Sharing and Dissemination** The software tools developed under NSF support will be released under an open-source license (to be determined) and available for download at the project website 6 months after the end of the project. Terms and conditions (e.g. authorship, acknowledgment, other compensation) of sharing will be negotiated by the PIs on a case by case basis. Given the nature of the data, there are no privacy, legal, regulatory, or ethical issues associated with data dissemination. As much as possible, results will be published in open access journals~~.~~ Tufts Provost's Open Access Fund provides monetary support for open access publication, and additional funds have been budgeted for publication (see <http://sites.tufts.edu/scholarlycommunication/?page_id=340>).