Open Science Guidelines / Policies

Allen Institute for Neural Dynamics

Sharing and distribution of data, tools, and knowledge is an imperative for AIND. We share the generosity of our founder and other sponsors to accelerate scientific discovery beyond AIND. One measure of our success is the value we and the larger scientific community derive from the scientific and technical resources we produce. Sharing data, tools, and resources maximizes overall scientific progress and facilitates reproducible science.

Sharing must happen in the context of community. We aspire to provide rapid access to our data, tools, and resources to facilitate their use. We want our scientific outputs to align with those of others, adhering to and developing community-based standards, to enhance collaboration throughout the field.

Our goal is to remove barriers to make our data, tools, and resources as accessible and usable as possible.

0. Principles

1. **Share in a community**. Actively engage with collaborators and colleagues in the field. Trust and seek community feedback. We want to do science together.
2. **Share intentionally**. Include metadata, documentation, and tutorials to facilitate use and re-use.
3. **Share freely**. Prioritize freedom of use over free cost but minimize cost.
4. **Share rapidly**. Even if imperfect.
5. **Share fairly**. Adhere to FAIR standards. Engage with community standards and domain-specific data repositories.

1. Discoveries

Discoveries will be described in preprints deposited to standard archives (bioRxiv, arXiv). In most cases, publication in a preprint archive will be accompanied by submission to a scientific journal. The cadence and granularity of dissemination will be determined on a case-by-case basis using our publication policies.

2. Data

Definitions

Data assets are logical units of data. They can be raw or derived data and can be a single file or an organized collection of files. Data assets are from a single experiment (e.g. single recording session, a single imaging acquisition).

Datasets are a collection of data assets that pertain to a specific project or publication. Data assets that are part of a dataset are likely to have undergone more extensive curation, quality control, and analysis than other data assets. As such, data assets can be labeled with metadata as belonging to a specific dataset – either at the time of acquisition or retroactively. A data asset can be part of multiple datasets.

ALLEN INSTITUTE for NEURAL DYNAMICS

Data Formats

Data assets should be stored in the standard data formats for data modalities and subfields. The INCF has resources on data standards that have been vetted, and we comply with these when appropriate. If a standard does not exist for a modality or does not meet our needs, we consult with local and community experts before designing our own.

We use NWB:N for neurophysiology (ephys and ophys) and OME formats for neuroimaging data. We prefer cloud-friendly variants of these standards (NWB-Zarr, OME-Zarr).

We push file format standardization as close to acquisition as possible. This helps minimize on-premise data transformations and processing time while giving us access to standardized tools. Exceptions may be necessary if formats are poorly optimized for the cloud, do not meet acquisition rate standards, or the cost of upgrading acquisition software is prohibitive. In these situations, we engage with our community partners to improve standards and, if possible, seek funding to upgrade acquisition software.

Metadata

All data assets must be accompanied by metadata that includes detailed information about how the data was acquired. Our metadata schemas are defined and maintained here. Metadata tracking procedures and data processing rely heavily on protocols being documented and maintained (see below).

At ingest, data should have metadata regarding:

- Data description (administrative information of the data asset)
- Subject (describes the animal subject used for the experiments, including background and genotype metadata, etc.)
- Procedures performed on the subject or tissue (including surgeries, injections, training, perfusion, etc.)
- Instrument of data acquisition (rig or microscope)
- Acquisition session (how the data is acquired)

As data is processed, additional metadata will track data processing steps performed and be associated with derived data assets.

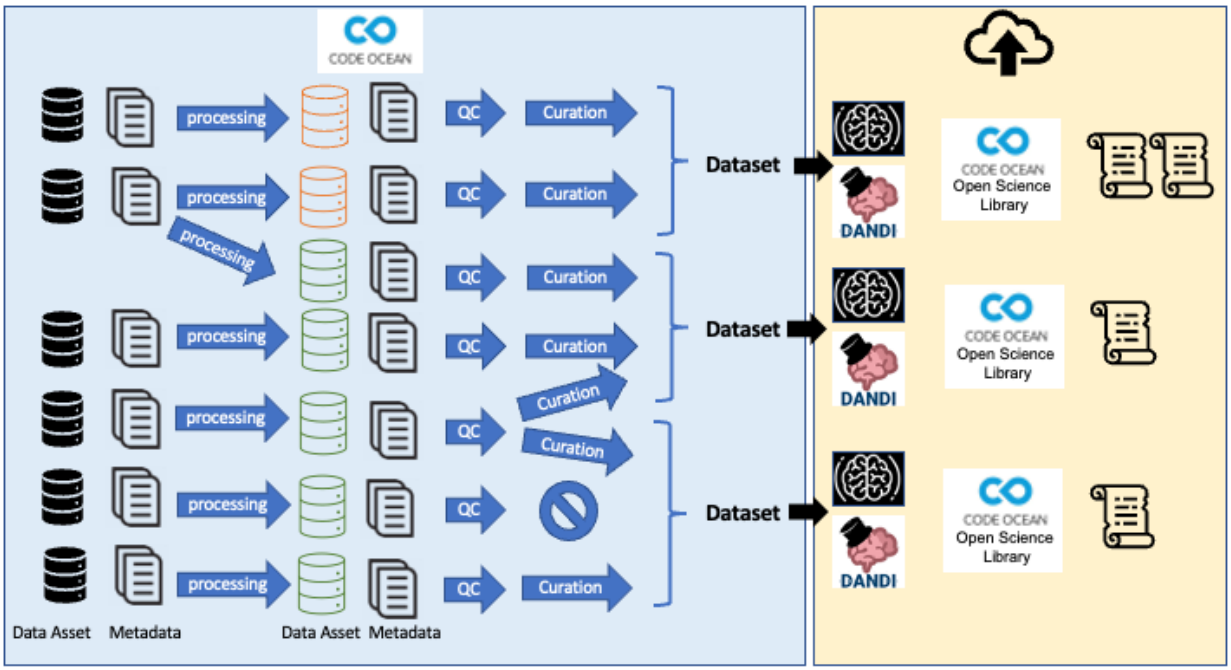
Finally, when a “dataset” is complete, there will be metadata describing that dataset as a whole.

Releasing Data

We release raw data, processed data, and analysis products along with metadata. Data might be part of a specific *dataset* or *project,* or it might be pilot data used for development.

Data assets are released by putting them in a public bucket in the cloud. Data will be compressed before it is uploaded to the cloud. In most cases, uncompressed raw data is deleted locally.

Some curated datasets may be shared in other forums beyond our public bucket on AWS. Specifically, datasets will be uploaded to appropriate BRAIN initiative data repositories, including DANDI for neurophysiology and BIL for imaging data. This sharing will likely be for curated datasets, usually linked with publications or resources.



We share data as soon as possible to minimize barriers to use and discovery. Deciding what data is worth releasing can be challenging and project-specific. We prefer to release data early with metadata indicating its status. This has the added benefit of encouraging us to use (and improve) the same analysis tools we offer to the rest of our community, rather than relying on custom internal solutions.

Public data must meet the following requirements:

- Data collection is complete. The data must have passed quality assurance, but various stages of quality control may still be ongoing.
- Metadata must contain all required fields defined in our metadata schemas. Metadata can be updated or corrected after initial release, with version control.
- Data must have a license allowing resharing and adaptation and requiring attribution (e.g. CC-BY-4.0).

Data will continue to be processed and analyzed when it is public, and new data assets will be derived from the public data. We use metadata to version control derived data assets (e.g. flags for the sanctioned spike sorting outputs).

There might be situations where we need to delay making data public rather than release it as soon as possible. These situations should be exceptional and need to be determined ahead of time and agreed upon by the team.

Publicizing Data

Data might be publicized when it is first released, but it is more likely to be publicized at a later date. This can include content on our website highlighting the data, press releases, social media, and other forms of publicity. Most likely, publicity will occur around publications and to highlight specific curated

datasets. However, publicity may also occur on a regular cadence, highlighting new data assets that have been added in the past time interval (e.g. every six months).

Public data also features regularly in ad hoc workshops to introduce scientists to our data platform. This includes the Summer Workshop on the Dynamic Brain, regular workshops with the UW CNC, conference workshops, and other events.

Data Citations

Data assets and datasets need to be associated with DOIs which will allow them to be citable. Data assets should have DOIs on upload to a public bucket. Datasets should have DOIs on publication. Our website needs to have clear instructions on how to cite the data.

3. Code

Standards for code and how we manage it follow directly from our Open Science principles.

Repositories

All software should be developed in open, public repositories unless third-party intellectual property or embargoes are involved. All software repositories should be part of the AllenNeuralDynamics GitHub organization.

Licenses

All new repositories should adopt permissive, OSI-approved licenses (MIT, BSD 2-Clause, BSD 3-Clause, or Apache v2.0). Pre-existing software in use should be licensed under the most permissive terms possible.

Languages

Software developers should use the right tool for the job, including the programming language. However:

- **Prefer open language ecosystems** (Python, R, C/C++, Java, Julia) **over proprietary ones** (Matlab). We want to develop software that is freely shareable with the community.
- **Prefer Python for scientific computing**. Agility and adoption increase the more we use common tools that are shared by our user community.

Dependencies

All tools developed at the institute will depend on a wide variety of external software libraries. The needs of an individual project will vary considerably, however:

- To simplify adoption, **prefer libraries that have OSI-approved, permissive licenses**.
- To simplify maintenance and installation, **prefer libraries that are actively maintained with up-to-date dependencies**.
- When changes in upstream libraries are needed, **prefer to contribute rather than forking or duplicating code**.

Standards

We aim for a high standard of code quality, but standards vary depending on the purpose of the code.

ALLEN INSTITUTE for NEURAL DYNAMICS

All code: All code should be in the AllenNeuralDynamics GitHub organization with a `README.md` indicating the level of support, a `LICENSE` file containing a permissive license (prefer MIT), and optionally a `requirements.txt` file. Template repository for general code.

Research and analysis code: Proof-of-concept software meant to demonstrate the feasibility of an approach. `README.md` should describe the data used and how to access it. Template repository for research and analysis code.

Publication code: Software written to generate results and figures for a publication. `README.md` should link to the publication and document how to access data, compute results, and generate figures. The repository should include a Docker file documenting the software environment used to produce results. The code should be peer-reviewed by at least one other qualified reviewer.

Tools and libraries: Software meant to be actively used and depended upon by others. We do not distinguish between tools meant for external vs. internal use. Code should be peer-reviewed by at least one other qualified reviewer. Template repository for tools and libraries.

Fully released tools and libraries must:

- have a DOI registered at zenodo.org
- have full*, automated test coverage
- be easily installable (via pip for Python) with semantic versioning
- have releases uploaded to standard package repositories (PyPi for Python)
- have user documentation uploaded to readthedocs.org
- have continuous integration running tests and building documentation (via Github Actions)
- have modern (< 1 year old) dependencies
- have a citation policy in `README.md`

* Full test coverage is required for release. Alpha or beta software *should* have full test coverage, but this is not required. Software in alpha/beta status should communicate this status in its version number and README.

Levels of Support

The following language should be used to describe the level of support guaranteed by software. It should be included in the `README.md` file for all software repositories.

- **Unsupported:** We are not currently supporting this code, but simply releasing it to the community AS IS but are not able to provide any guarantees of support. The community is welcome to submit issues, but you should not expect an active response.
- **Occasional updates:** We are planning on occasionally updating this tool with no fixed schedule. Community involvement is encouraged through both issues and pull requests.
- **Supported:** We are releasing this code to the public as a tool we expect others to use. Issues are welcomed and we expect to address them promptly, pull requests will be vetted by our staff before inclusion.

ALLEN INSTITUTE for NEURAL DYNAMICS

4. Hardware

Researchers need to be able to replicate the physical rigs we use for data collection. Sharing hardware presents several challenges that do not apply to software:

- There are fewer standards for hardware documentation, file formats, and repository structures.
- Many labs lack the ability to recreate custom hardware components. It can make sense to organize reproduction via a third party (e.g. a company), rather than relying on individual labs to build the tools themselves. Quantity discounts can be an additional bonus.

Because of these reasons, there is not a "one-size-fits-all" solution for sharing hardware. Moreover, it may not make sense to make all of our hardware open source (see 0. Principles).

Electronics

Printed circuit board (PCB) designs are shared as Git repositories on the AIND GitHub. The repository should contain the schematics, board layouts, and bill of materials (BOM) as KiCad project files, which are version controlled. Each time we manufacture a new board revision, we should create a GitHub release containing the incremented revision number (use semantic versioning), schematic PDF, and the Gerber and pick-and-place files needed to manufacture the boards. The repo should contain a README describing the board's intended use, and the interfaces it presents (SPI, USB, GPIO pinout, etc.). It should also contain a LICENSE file compatible with the Open Source Hardware Definition; this includes (but is not limited to):

- the TAPR license (copyleft)
- the Solderpad license (permissive)
- the CERN Open Hardware License (available in both permissive and restrictive/copyleft versions).

For an example open-hardware repository conforming to these standards, see mis-focus-controller.

Firmware

Some electronic circuits we develop will be controlled by a microcontroller, FPGA, or other embedded device on which we will need to run firmware, defined as embedded software running without an operating system. Sharing firmware is similar in many ways to sharing application-level software but with some important differences. Firmware is typically specific to a particular hardware architecture, and often to a specific chip. It is typically compiled into a binary, then "flashed" or "burned" onto the hardware, where it then runs autonomously on power-up. Thus, the firmware is often found running "in the wild" without clear provenance to the source code that produced it. It also depends on a (typically large) stack of dependencies, known as the Hardware Abstraction Layer (HAL), which also needs to be shared for transparency.

Firmware at AIND should be shared in a Git repository on the AIND GitHub. The repository should contain the source code (prefer C), HAL (prefer open-source libraries included as git submodules), and compiling instructions (prefer Makefile), as well as the instructions to flash onto the hardware. These instructions should be included in a README file, which also describes the basic usage of the firmware. There should be a clear way to query the firmware version currently running on the hardware, either

through a serial interface or debugging tool. The repository should be licensed through an OSHWA-compatible license; this may be one of the hardware licenses listed above or a software license such as MIT or GPL.

Assemblies including CAD files

For hardware we want to develop collaboratively with the community, it's important to have version-controlled design files. This is non-trivial due to the binary and proprietary nature of most of the file formats we would use. This can potentially be solved by providing remote access to the MPE vault, but we will need to test the feasibility of this approach.

Simple hardware designs can be distributed as STEP or SVG files in a GitHub repository (example here), along with detailed manufacturing instructions. In some cases, it may make sense to partner with a company that can handle the dissemination of these parts, in order to facilitate access by a wider audience.

5. Reagents

Reagents will be deposited in standard repositories if available and as soon as they are characterized (i.e. generally before publication). Exception may have to be made if third-party intellectual property or other embargoes apply.

Mice – JAX

Plasmids – Addgene

6. Protocols

Protocols are procedures or methods related to the implementation of an experiment. Protocols include a list of required tools and materials, checklists, step-by-step instructions, and didactic material such as photos, diagrams, and video clips. High-quality protocols are a key step in documenting our workflows and are indispensable for reproducible science. Sharing high-quality protocols in a public & permanent repository allows our work to be reproduced and built on. Shared protocols also greatly simplify the writing of research papers. Protocols are also a citable output from our research.

We adopt Protocols.io.

- Each protocol starts out private as part of an AIND group (i.e. visible to all AIND staff and collaborators, but not others).
- Protocols have to be reviewed by at least two scientists.
- Protocols must have a permanent DOI and is thus citable.
- Protocols and their DOI must be created before data acquisition. They need to be part of the metadata during acquisition.
- Protocols are published (made open with a DOI) at the time of acquisition.

Protocols are archived (CLOCKSS) for long-term knowledge preservation.

7. Website

We will have a website for our open science / open technology projects which routes to our GitHub, Zenodo, Code Ocean, etc. profiles. Details of this site will be captured in a separate document.