DATA MANAGEMENT AND SHARING PLAN

Franco Pestilli, University of Texas at Austin

Element 1: Data Type

A. Types and amount of scientific data expected to be generated in the project:

Although the project will not generate new data, we estimate that the proposed archive will store approximately 250TB within its first year. After the first year, we estimate up to 500TB will be deposited annually. These data will be decentralized such that up to 300 TB are stored at Texas Advanced Computing Center (TACC). We will utilize commercial cloud storage through Microsoft Azure and with an unlimited amount of storage at Amazon Web Service (AWS) through the Amazon Open Data Program. The types of data that will be stored include actigraph, human videos, primate videos, rodents videos, eye tracking, trials and events, sensory stimuli, virtual reality, eye tracking, trials and events, sensory stimuli, virtual reality, psychophysics, human pose estimation.

B. Scientific data that will be preserved and shared, and the rationale for doing so:

BEHIVE will provide a secure, cloud-based, and user-friendly data ecosystem for storing, tracking, analyzing, and sharing multimodal behavioral data. It will support diverse data modalities used in humans and non-human animal models primates (macaque and marmoset), rodents (rats, mice, and hamsters) and birds, such as videographic, audiographic, electrophysiologic, imaging, eye movements, vocalizations aligned with corresponding stimuli. By aligning with NIH data management and sharing policies, BEHIVE will facilitate the secure, cloud-based, and user-friendly storage and analysis of behavioral data.

C. Metadata, other relevant data, and associated documentation:

The proposed BEHIVE will be designed to work seamlessly with computational platforms such as Brainlife, NEMAR, CBRAIN, and DataJoint. Integrations will also extend to the existing data archives such as DABI, DANDI, BossDB, and OpenNeuro. BEHIVE DataType metadata records will be synchronized and harmonized from the metadata records for the datasets and files (assets in DANDI archive terms) which potentially would follow their own schema. The effort will be made to converge on use of the same set of ontologies and metadata dictionaries to minimize the necessary amount of metadata harmonization, and instead benefit from centralized ontologies, e.g. as the one being developed within INCF Behavioral tasks and paradigms ontology. BEHIVE DataTypes will be uniformly and sufficiently documented and readily available on the platform, with references to underlying used Ontologies where applicable. Detailed user-oriented documentation and walkthroughs will be provided on various aspects of metadata entry and editing, and released under a permissive Creative Commons license.

Element 2: Related Tools, Software and/or Code:

Support from the BRAINLIFE and DataJoint teams will enable the deployment of cutting-edge tools for data analysis and visualization, enabling the application of machine learning and artificial intelligence. The proposed BEHIVE will be designed to work seamlessly with computational platforms such as BRAINLIFE, NEMAR, CBRAIN, and DataJoint. Integrations will also extend to the existing data archives such as DABI, DANDI, BossDB, and OpenNeuro. These expansive integrations will rely on a decentralized data tracking technology such as DataLad. All of these tools, software, analysis code, and data archives are available online, and most of them (such as DANDI, DataLad, etc) are Free and Open Source Software released under permissive OSI-compliant licenses. The code developed in the scope of the BEHIVE project will also be released under OSI-compliant licenses as appropriate (primarily Apache-2). Additional visualization tools that collaborators will help contribute to the proposal include but are not limited to:

Toolbox	Point of contact	Grant Number	Function Table 1
NBviewer	Jupyter Project	n.a.	Visualization in Jupyter
NIVue	Chris Rorden		Brain
MoBILAB	Scott Makeig	R24-MH120037	EEG & body synchronization
Video Segment Analysis (no tool)	Chen Yu	R01-HD093792	Segmentation of video images
easyeyes	Denis Pelli	n.a.	Web measurement eye movement
Naturalistic Human Behavior	Kate Bonner, Jon Mathis	n.a.	Tracking human movement
PsychToolbox			Measurement of visual and auditory stimulation
PsychJS	Josh De Leuw	n.a.	Web-based cognitive experiments
Psiturk	Todd Gureckis		Web-based, Amazon Turk Behavioral experiments
MGL	Justin Gardner	n.a.	MatLab psychophysics & behavior
PsychoPy	Jonathan Peirce	n.a.	Measurement of visual and auditory behavior
OpenMonkeyStudio	Jan Zimmerman	R01-MH128177	Automated markerless pose estimation
colliga.io	Adela Timmons	R44-MH123368	Mobile behavioral research studies

Element 3: Standards:

We will consider all standards that our advisory board and collaborators will suggest. Currently, we are listing some of the most prominent project standards for behavioral data that we will be considering, see Table XX. Dr. Pestilli is collaborator of the Hierarchical Event Descriptor (HED) and Dr. Halchenko of Neurodata Without Borders (NWB). The standards all have published documentation that can be accessed on public websites or GitHub. The majority of these standards have been developed under BRAIN Initiative awards. We have enlisted collaborators that will help contribute the standards to the proposal (see LoCs).

Standard/Toolbox	Point of contact	Grant Number	Function Table 2
BIDS & BIDS Derivatives	Russ Poldrack, Franco Pestilli, Dora Hermes	R24-MH114705, OAC-1760950 R01-MH126699	A specification on a way to organize and describe your neuroimaging and behavioral data (BEP007 Task Data; BEP020 Eye Tracking; BEP036 Phenotypic Data)
BEADL & BAABL	Adam Kepecs	RF1-MH120034	A Universal Framework for Describing Behavioral Tasks
NWB & ndx-events and other extensions	Oliver Rubel, Ben Dichter	R24-MH116922 U24-NS120057	A standardized data structure designed to facilitate the organization, sharing, and analysis of neurophysiological and associated behavioral data.
HED (Hierarchical Event Descriptor)	Scott Makeig, Kay Robbins	RF1-MH126700	A system to describe events in human behavioral tasks data and to synchronize the data with neuroimaging & EEG
Psych-DS	Melissa Kline	n.a.	A specification for psychological data following BIDS
Squirrel Data Model NiDB	Gregory Book	n.a.	Data format that allows sharing information necessary to recreate experimental results.

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:

The proposed archive (BEHIVE) will store the behavioral data and metadata. Similarly to the procedures established by DANDI archive, open metadata dumps will also be archived on AWS S3 to guarantee longevity of the metadata to transcend possible life-span of the BEHIVE effort itself.

B. How scientific data will be findable and identifiable:

BEHIVE will assign persistent identifiers to published data sets. In response to the NOSI BEHIVE will require mature technology from the team to assign persistent identifiers to datasets and associated metadata. The team has experience with assigning Digital Object Identifiers (DOI) to published datasets. Indeed, both DABI, DANDI and BRAINLIFE assign DOIs to published data sets (see Avesani et al., 2019). BRAINLIFE and DANDI use DOIs emitted by the DataCite consortium. We will reuse the same approach and use DOIs for BEHIVE. Importantly, BRAINLIFE developed an innovative approach and assigns a single DOI to a published record, a bundle containing not just data but data objects, data processing applications and Jupyter Notebooks used for any data post-processing or visualization, importantly a provenance information record is also associated with the data record providing tracking, versioning and synchronization metadata. By reusing this approach, BEHIVE will allow researchers to share their data and metadata with groups of collaborators and the scientific community. Within the archive we will establish versatile (basic keyword and faceted search) mechanisms based on the DataType records and metadata records for the data residing in the other archives. To expose BEHIVE datasets to wider community and generic data search engines such as Google Dataset Search, similarly to how done by OpenNeuro, we will expose datasets metadata via JSON-LD records using schema[dot]org context.

C. When and how long the scientific data will be made available:

The BEHIVE archive will implement a graded data release process. This process involves releasing data in stages or tiers, allowing for controlled and responsible data sharing. The release process may involve different levels of access restrictions based on data sensitivity, ethical considerations, or legal requirements. Data may be initially released to a restricted group of researchers or collaborators and gradually made available to a wider audience as appropriate.
By adhering to the FAIR principles and implementing a graded data release process, the BEHIVE archive aims to strike a balance between data accessibility and data stewardship. It ensures that data can be discovered, accessed, and reused effectively while considering privacy, security, and ethical considerations associated with data sharing. Currently we have no temporal limit established for the data to be stored in the archive. So far, AWS Open Data program in-kind support of OpenNeuro and DANDI archives did not mandate any termination date. To guarantee long term availability, we will also join efforts of those archives to research alternative data storage platforms and approaches, e.g. explored by the DANDI FileCoin “digital currency” approach allowing to efficiently use next generation distributed storage model of the Interplanetary File System (IPFS). As demand arises, we will also consider establishing multi-tier data storage with migrating some rarely accessed data to cheaper storage models such as AWS Glacier.

Element 5: Access, Distribution, or Reuse Considerations

A. Factors affecting subsequent access, distribution, or reuse of scientific data:

BEHIVE aims to be a comprehensive and accessible platform for storing, integrating, and sharing diverse behavioral datasets. BEHIVE will align with NIH FAIR data access policies. BEHIVE will ensure that data can be discovered, accessed, and reused effectively while considering privacy, security, and ethical considerations associated with data sharing.

Reusability: The BEHIVE archive will focus on maximizing the reusability of data. It will provide clear and comprehensive documentation on data collection methods, experimental protocols, and data processing pipelines. Metadata associated with the datasets will include information on data provenance, quality, and any transformations applied. This will enhance the reproducibility and interpretability of the data and enable researchers to reuse it for various analyses and investigations.

B. Whether access to scientific data will be controlled:

Accessibility: The BEHIVE archive will prioritize making data accessible to users. It will provide multiple access options such as web-based graphical user interfaces (GUIs), command line interfaces (CLIs), and application programming interfaces (APIs). Access controls and permissions will be implemented to ensure that data is appropriately shared while respecting privacy and data sensitivity.

C. Protections for privacy, rights, and confidentiality of human research participants:

BEHIVE will store multiple types of data that will represent a wide spectrum of behaviors from humans, non-human primates (macaque and marmoset), rodents (rats and mice) and birds. Some of the files will require protection of study participants as well as of the user submitting the data. This is for example, the case of video data of individual humans and families interacting in their homes, or images of birds, rodents and primates interacting while being recorded. Because of the sensitive nature of some of these datasets we will implement a proficient approach to FAIR data access and management. There are several privacy concerns to consider for the data. Video data captures sensitive information and personal interactions, so it's crucial to handle the data responsibly and protect the privacy of the participants. Here are some key considerations and technological solutions to address these concerns:

Anonymization and De-identification: To protect participants' identities, we will require data depositors to comply with best practices for each type of data to perform pseudo-anonymization and de-identification. This helps minimize the risk of re-identification.

Secure Data Storage: We will implement robust security measures for data storage. Examples include encryption of data at rest and in transit to prevent unauthorized access. BEHIVE will utilize secure cloud storage or on-premises infrastructure with strict access controls and monitoring.

Access Control: We will implement strong access control mechanisms to ensure that only authorized researchers can access data, such as role-based access control (RBAC) to define different levels of access based on researchers' roles and responsibilities and use the industry-standard OAUTH to authenticate users (as in BrainLife).

Data Sharing Protocols: We will develop streamlined data sharing protocols that define the conditions and procedures for data access. Researchers requesting access would provide appropriate justifications and comply with ethical guidelines and legal requirements. The DAC will review the most critical datasets and access requests when the privacy of the individuals is key, a threshold that will be formally operationalized in collaboration with the Advisory Board.

Data Usage Agreements (DUA): BEHIVE will provide researchers sharing data DUA templates (to outline their responsibilities, including data handling, privacy preservation, and adherence to ethical guidelines). The BEHIVE GUI will serve DUAs to users accessing data. Ultimately, data owners will be responsible for enforcing compliance and auditing. Help and best practices for enforcing DUAs will be part of the AIM 3 community knowledge transfer (e.g., during office hours, within the Slack community).

Element 6: Oversight of Data Management and Sharing:

Ethical Oversight: Establish an ethical review board or seek institutional review board (IRB) approval to ensure that your research protocol and privacy protection measures align with ethical standards and legal requirements.

An experienced Advisory Board (AB) will help the team manage priorities and implement best practices in the field with special emphasis on toolboxes, data to be handled by the archive.

OUTPUTS

Title	Type	Access
brainlife.io	software	Open Access