The Neurodata Without Borders ecosystem for neurophysiological data science

Oliver Rübel[1*$], Andrew Tritt[1*], Ryan Ly[1*], Benjamin K. Dichter[2*], Satrajit Ghosh[3,4*], Lawrence Niu[5], Ivan Soltesz[6], Karel Svoboda[7,8], Loren Frank[8,9,10], Kristofer E. Bouchard[1,9,11,12,13*$]

***:** equal contributors; $: contact authors

Key words: neurophysiology, data ecosystem, data language, data standard, archive, FAIR data

Affiliations:

1. Computational Research Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
2. CatalystNeuro, Benicia, CA, USA
3. McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA
4. Department of Otolaryngology - Head and Neck Surgery, Harvard Medical School, Boston, MA, USA
5. Vidrio Technologies
6. Department of Neurosurgery, Stanford University, Stanford, CA, USA,
7. Janelia Research Campus
8. Howard Hughes Medical Institute
9. Kavli Institute for Fundamental Neuroscience, San Francisco, CA, USA
10. Departments of Physiology and Psychiatry University of California, San Francisco, CA, USA
11. Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
12. Helen Wills Neuroscience Institute and Redwood Center for Theoretical Neuroscience, University of California, Berkeley, CA, USA
13. Weill Neurohub

Abstract: The neurophysiology of cells and tissues are monitored electrophysiologically and optically in diverse experiments and species, ranging from flies to humans. Understanding the brain requires integration of data across this diversity, and thus these data must be findable, accessible, interoperable, and reusable (FAIR). This requires a standard language for data and metadata that can coevolve with neuroscience. We describe design and implementation principles for a language for neurophysiology data. Our software (Neurodata Without Borders, NWB) defines and modularizes the interdependent, yet separable, components of a data language. We demonstrate NWB's impact through unified description of neurophysiology data across diverse modalities and species. NWB exists in an ecosystem which includes data management, analysis, visualization, and archive tools. Thus, the NWB data language enables reproduction, interchange, and reuse of diverse neurophysiology data. More broadly, the design principles of NWB are generally applicable to enhance discovery across biology through data FAIRness.

Introduction

The immense diversity of life on Earth[1] has always provided both inspiration and insight for biologists. For example, in neuroscience, the functioning of the brain is studied in species ranging from flies, to mice, to humans (**Fig. 1a**)[2]. Because brains evolved to produce a plethora of behaviors that advance organismal survival, neuroscientists monitor brain activity while subjects are engaged in diverse tasks (**Fig. 1a**). On one hand, creating a coherent model of how the brain works will require synthesizing data generated by these heterogeneous experiments. On the other hand, the extreme heterogeneity of neurophysiological experiments impedes the integration, reproduction, interchange, and reuse of diverse neurophysiology data. A standardized language for neurophysiology data and metadata (i.e., a data language) is required to enable neuroscientists to effectively describe and communicate about their experiments, and thus share the data.

The extreme heterogeneity of neurophysiology experiments is exemplified in **Figure 1**. Diverse experiments allow neuroscientists to study how brain activity supports sensation, perception, cognition, and action. Tasks range from running on balls or treadmills (e.g., pictures, **Fig. 1i**)[3], memory-guided navigation of mazes (**Fig. 1ii)**[4], production of speech (**Fig. 1iii)**[5], and memory formation (**Fig. 1iv**). The use of different species in neuroscience is driven, in part, by the applicability of specific neurophysiological recording techniques (**Fig. 1b**). For example, the availability of genetically modified mice makes this species ideal to monitor the activity of genetically defined neurons using protein calcium sensors (e.g., with GCaMP;optophysiology, 'o-phys') (**Fig. 1bi,ci**). Concomitantly, electrophysiology probes ('e-phys') with large numbers of electrodes enable monitoring the activity of multiple single neurons at millisecond resolution from different brain regions (e.g., medial prefrontal cortex, 'mPFC'; ventral striatum, 'v. Striatum'; orbital frontal cortex, 'OFC') in freely behaving rats (**Fig. 1bii,cii**)[4]. Additionally, in human epilepsy patients, arrays of electrodes on the cortical surface (i.e., electrocorticography, ECoG) allows direct electrical recording of mesoscale neural activity at high-temporal resolution from entire brain regions and multiple brain areas (e.g., speech sensorimotor cortex 'SMC'; **Fig. 1biii, ciii**)[5]. To characterize the functioning of single neurons, scientists measure membrane potentials (ic-ephys), e.g, via patch clamp recordings. As a final example, to study the detailed workings of complete

neural circuits, supercomputers are used for biophysically detailed simulation of the intracellular membrane potentials of a large variety of neurons organized in complex networks[6,7] (**Fig. 1biv,civ**). See also **Supplementary Material 1** for description of intracellular electrophysiology data. Although this heterogeneity is most evident across labs, it is present in a reduced form within single labs; lab members can use different techniques in custom experiments to address specific hypotheses. As such, even within the same laboratory, it is not uncommon for data and metadata to be described in an idiosyncratic way by the individual who collected the data, making the data nearly useless when that individual leaves the lab.

Figure 1

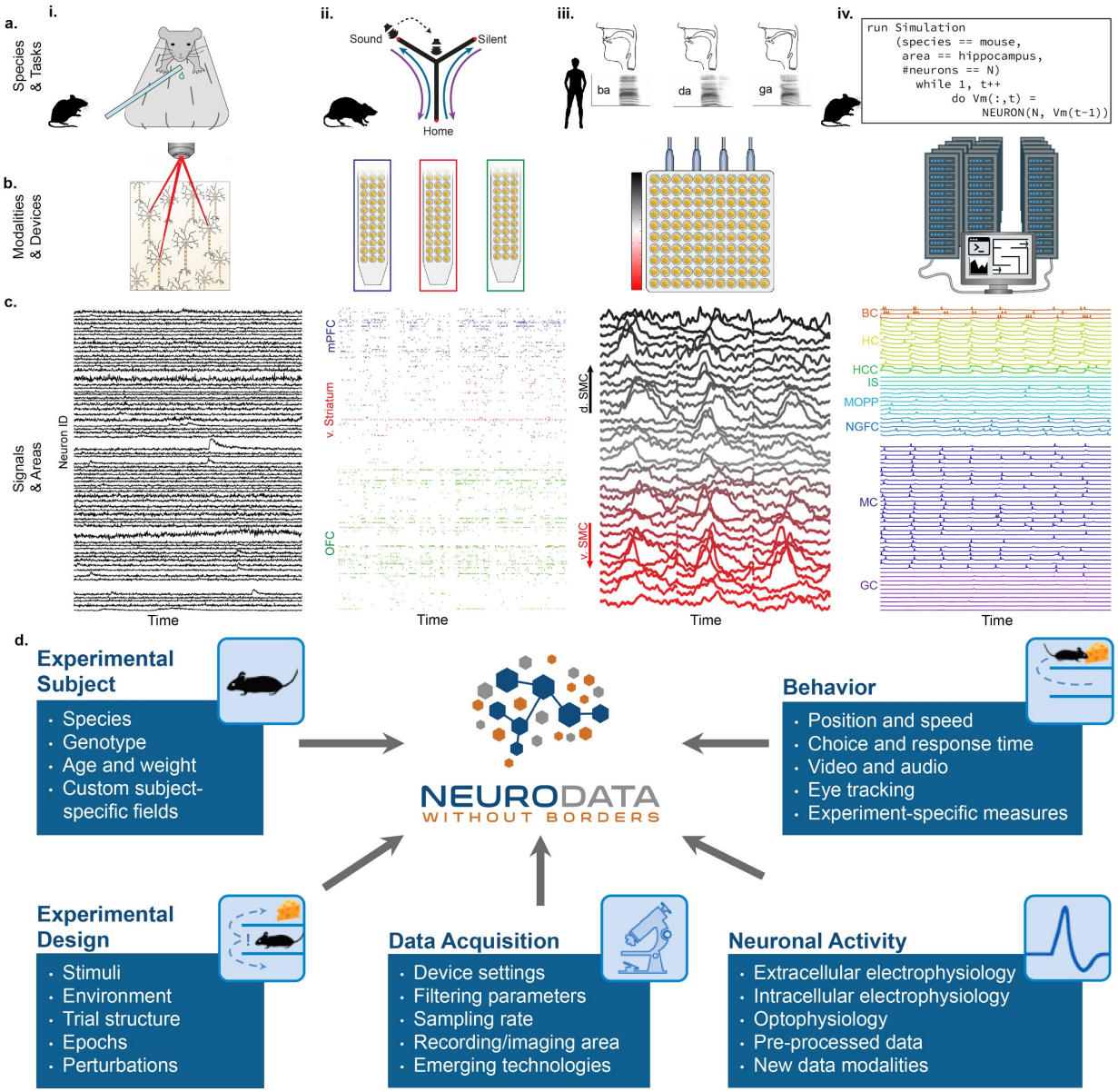


Figure 1: NWB addresses the massive diversity of neurophysiology data and metadata

a. Diversity of experimental systems: species and tasks. (**i**). mice performing a visual discrimination task; (**ii**) rats performing a memory-guided navigation task; (**iii**) humans speaking consonant-vowel syllables; (**iv**) biophysically detailed simulations of mouse hippocampus during memory formation. The corresponding acquisition modalities and signals are shown in the corresponding columns in figure **b** and **c**.

b. Diversity of data modalities and acquisition devices: (**i**) optophysiological Ca^{2+} imaging with 2-photon microscope; (**ii**) intra-cortical extracellular electrophysiological recordings with polytrodes in multiple brain areas (indicated by color, see **cii**); (**iii**) cortical surface electrophysiology recordings with electrocorticography grids; (**iv**) high-performance computing systems for large-scale, biophysically detailed simulations of large neural networks.

c. Diversity of signals and areas: (**i**) Ca^{2+} signals as a function of time from visually identified individual neurons in primary visual cortex (V1)(Mallory et al., 2021); (**ii**) spike-raster (each

tick demarcates the time of an action potential) from simultaneously recorded putative single-units after spike-sorting of extracellular signals from medial prefrontal cortex (mPFC; blue), ventral striatum (v. Striatum, red), and orbital frontal cortex (OFC, green)(color corresponds to **b.ii**)(Kastner et al., 2020); (**iii**) high-gamma band activity from electrodes over the speech sensorimotor cortex (SMC), with dorsal-ventral distance from Sylvian fissure color coded red-to-black (color corresponds to **b.iii**) (Bouchard et al., 2013); (**iv**) simulated intra-cellular membrane potentials from different cell-types from large-scale biophysical simulation of the hippocampus (BC, Basket Cell; HC, Hilar Interneuron (with axon associated with the) Perforant Path; HCC, Hilar Interneuron (with axon associated with the) Commissural/Associational Path; IS, Interneuron-Specific Interneuron; MCPP, medial Perforant Path; NGFC, neurogliaform cell; MC, mossy cell; GC, granule cell](Raikov and Soltesz, unpublished data).

d. Neurodata Without Borders (NWB) provides a robust, extensible, and maintainable software ecosystem for standardized description, storage, and sharing of the diversity of experimental subjects, behaviors, experimental designs, data acquisition systems, and measures of neural activity exemplified in **a - c**. In addition to the examples in **i-iv**, another important application area for NWB is intracellular electrophysiology data (see Supplement Material 1).

Scientific data must be thought of in the context of the entire data life-cycle, which spans planning, acquisition, processing, and analysis to publication and reuse[8]. Across species and tasks, different acquisition technologies measure different neurophysiological quantities from multiple spatial locations over time. Thus, the numerical data itself can commonly be described in the form of space-by-time matrices, the storage of which has been optimized (for space and rapid access) by computer scientists for decades. It is the immense diversity of metadata required to turn those numbers into knowledge that presents the outstanding challenge.

As other fields of science, such as climate science[9], astrophysics[10], and high-energy physics[11] have demonstrated, community-generated standards for data and metadata are a critical step in creating robust data and analysis ecosystems, as well as enabling collaboration and reuse of data across laboratories. This issue is particularly relevant in the current age of massive neuroscientific data sets generated by emerging technologies from the US BRAIN Initiative, Human Brain Project, and other brain research initiatives worldwide. Advanced data processing, machine learning, and artificial intelligence algorithms are required to make discoveries based on such massive volumes of data[12,13,14]. There have been previous efforts to standardize neurophysiology data, such as NWB(v1.0) and NIX[15,16]. However, these previous efforts have not gained wide-spread adoption by the diverse neurophysiology community.

Why have these previous efforts not been successful? We argue that there are several challenges that have not yet been successfully addressed. Conceptually, the complexity of the problem necessitates an interdisciplinary approach of neuroscientists, data and computer scientists, and scientific software engineers to identify and disentangle the components of the solution. Technologically, the software infrastructure instantiating the standard must integrate the separable components of user-facing interfaces (i.e., Application Programming Interfaces, APIs), data modeling, standard specification, data translation, and storage format. This must be done while maintaining sustainability, reliability, stability, and ease of use for the neurophysiologist. Furthermore, because all science is advanced by both development of new acquisition techniques and creation of ingenious experimental designs, mechanisms for extending the standard to unforeseen data and metadata are essential. Sociologically, the neuroscientific community must accept and adopt the standard, requiring coordinated community engagement, software development, and governance. Together, these challenges and requirements necessitate a conceptual departure from the traditional notion of a static, monolithic data standard. That is, we need a "language" to enable precise communication about data that can co-evolve with the needs of the community.

We created the Neurodata Without Borders (NWB) data language (i.e., a standardized language for describing data) for neurophysiology to address the challenges described above. NWB(v2) accommodates the massive heterogeneity and evolution of neurophysiology data and metadata in a unified framework through the development of a novel data language that can co-evolve with neurophysiology experiments. Through extensive and coordinated efforts in community engagement, software development, and interdisciplinary governance, NWB is now being utilized by more than 36 labs and research organizations. Across these groups, NWB is used for all neurophysiology data modalities collected from species ranging from flies to humans during diverse tasks. This generality was enabled by the development of a robust, extensible, and sustainable software architecture based on our Hierarchical Data Modeling Framework (HDMF)[17]. NWB is integrated with state-of-the-art analysis tools to provide a unified storage platform throughout the data life-cycle, as we demonstrate through the analysis, visualization, and re-storage of diverse neurophysiological data and processing results. To facilitate new

experimental paradigms, we developed methods for the creation and sharing of NWB Extensions that permits the NWB data language to co-evolve with the needs of the community. Finally, NWB is foundational for the Distributed Archives for Neurophysiology Data Integration (DANDI) data repository to enable collaborative data sharing and analysis. Together, the capabilities of NWB provide the basis for a community-based neurophysiology data ecosystem. The processes and principles we utilized to create NWB provide an exemplar for biological data ecosystems more broadly.

Results

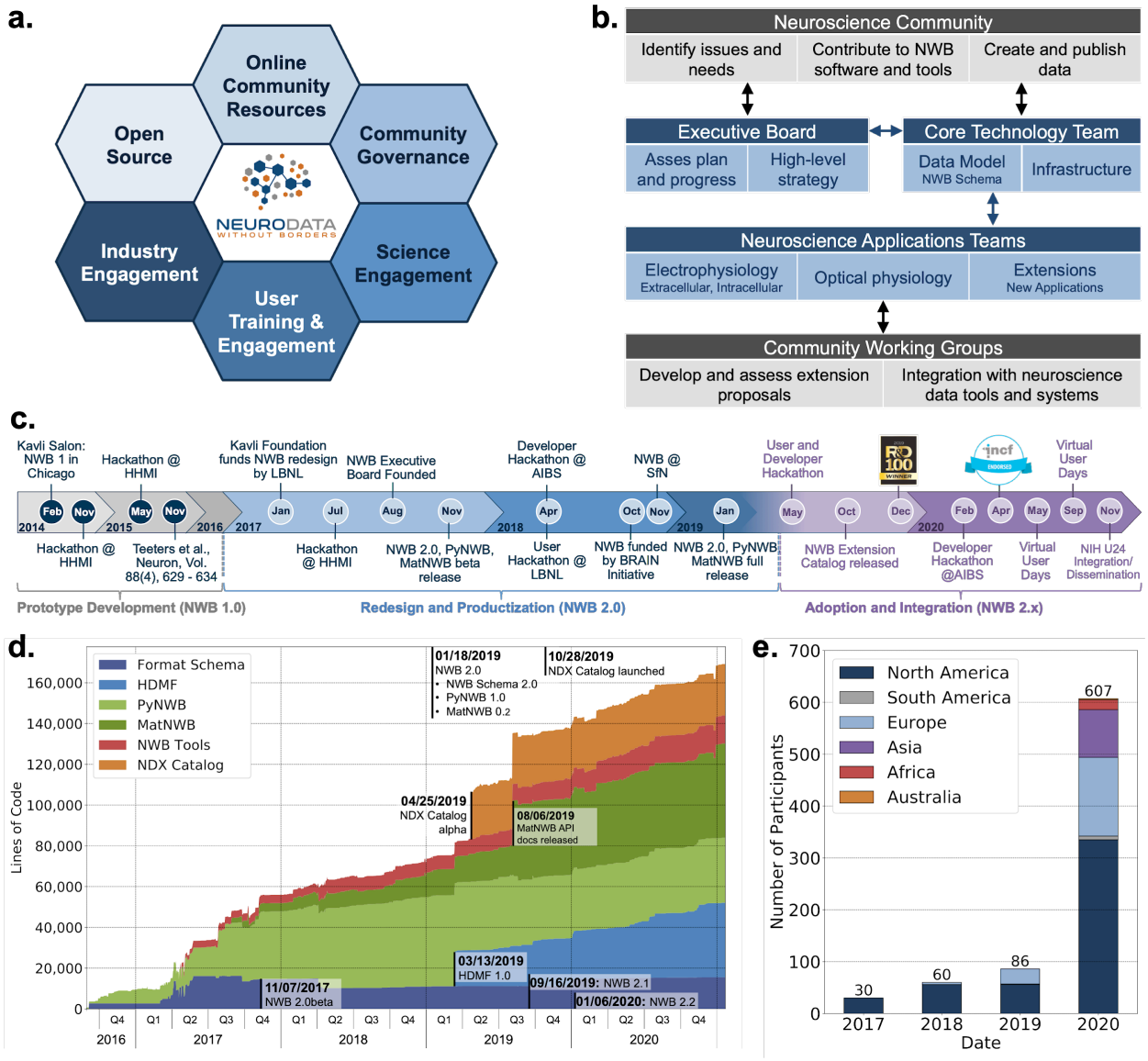


Figure 2. Coordinated Community Engagement, Governance, and Development of NWB

- **a.** NWB is open source with all software and documents available online via GitHub and the nwb.org website. NWB provides a broad range of online community resources, e.g., Slack, GitHub, mailing list, Twitter, to facilitate interaction with the community and a broad set of online documentation and tutorials. NWB uses an open governance model that is transparent to the community. Broad engagements with industry partners (e.g., Vathes, Kitware, MathWorks, Vidrio, CatalystNeuro, etc.) and targeted science engagements with both neuroscience labs and tool developers, help sustain and grow the NWB ecosystem. Broad user training and engagement activities, e.g., via hackathons, virtual training, tutorials at conferences, or online training resources, aim at facilitating adoption and growing the NWB community knowledge base.
- **b.** Organizational structure of NWB showing the main bodies of the NWB team (blue boxes) and the community (gray boxes), their roles (light blue/gray boxes), and typical interactions (arrows).
- **c.** The timeline of the NWB project to date can be roughly divided into three main phases. The initial NWB pilot project (2014-2015) resulted in the creation of the first NWB 1.0 prototype data

standard. The NWB 2.0 effort then focused on facilitating use and long-term sustainability by redesigning and productizing the data standard and developing a sustainable software strategy and governance structure for NWB (2017-2019). The release of NWB 2.0 in Jan. 2019 marked the beginning of the third main phase of the project, focused on adoption and integration of NWB with neuroscience tools and labs, maintenance, and continued evolution and refinement of the data standard.

d. Overview of the growth of core NWB 2.x software in lines of code over time.

e. Number of participants at NWB outreach and training events over time. In the count we considered only the NWB hackathons and User Days (see **c**.), the 2019 and 2020 NWB tutorial at Cosyne, and 2019 training at the OpenSourceBrain workshop (i.e., not including attendees at presentations at conferences, e.g., SfN).

Coordinated Community Engagement, Governance, and Development of NWB

There is great diversity of neurophysiology data and heterogeneity of application uses of data standards, ranging from data acquisition, processing and analysis, data management and integration, to data sharing, archiving and publication. This leads to a large diversity of stakeholders with vested interests and diverse use cases. Broad engagement and outreach with the community are central to achieve acceptance and adoption of NWB and to ensure that NWB meets user needs. Thus, development of scientific data languages is as much a sociological challenge as it is a technological challenge. To address this challenge, NWB has adopted a modern open source software strategy (**Fig. 2a**) with community resources and governance, and diverse engagement activities.

Execution of this strategy requires coordinating efforts across stakeholders, use cases, and standard technologies, prioritization of software developments, and resolution of potential conflicts. Such coordination necessitates a governance structure reflecting the values of the project (**Fig. 2b**). The NWB executive board consists of leading experimental and computational neuroscientists from the community, and serves as the steering committee for developing the long-term vision and strategy, and assess development plans and progress. The NWB Executive Board was established in 2007 as an independent body from the technical team and PIs of NWB grants. The NWB Core Technology Team leads and coordinates the development of the NWB data language and software infrastructure to ensure timely response to user issues and ensure quality, stability, and consistency of NWB technologies. The technology team reports regularly to and coordinates with the Executive Board. Targeted Neuroscience Application Teams consisting of expert users and core developers lead engagement with the diverse

neuroscience data areas in electrophysiology, optical physiology, and emerging applications. These application teams are responsible for developing extensions to the data standard, new features, and technology integration together with Community Working Groups. The working groups allow for an agile, community-driven development and evaluation of standard extensions and technologies, and allow users to directly engage with the evolution of NWB. The broader neuroscience community further contributes to NWB via issue tickets, contributions to NWB software and documentation, and by creating and publishing data in NWB. This governance structure emphasizes and seeks to balance stability of NWB technologies to ensure reliable production use, direct engagement with the community to ensure that NWB meets stakeholder needs, and agile response to issues and emerging technologies.

These values are also reflected in the overall timeline of the NWB project (**Fig. 2c**). The NWB 1.0 prototype focused on evaluation of existing technologies and community needs and development of a draft data standard. NWB 2.0 built upon version 1.0, enhancing adoption and software maintainability and extensibility by adding a clear governance structure, a complete redesign of the software, and integrating with neurophysiology tools. The NWB 2.0 project focused on the redesign and productization of NWB, emphasizing the creation of a sustainable software architecture, reliable data standard, and software ready for use. By analogy to microscopes, NWB 1.0 is like the first candle powered light microscope, while NWB 2.0 is a modern 2-photon microscope. Indeed, NWB has been recognized for its technological significance by the R&D 100 awards and NWB is the only neurophysiology data standard endorsed by both the International Neuroinformatics Coordinating Facility (INCF) and the US BRAIN Initiative. Following the first full release of NWB 2.0 in January 2019, the emphasis then moved to adoption and integration to grow a vibrant community and software ecosystem as well as maintenance and continued refinement of the data standard to ensure reliability and address community needs.

Together, these technical and community engagement efforts have resulted in a vibrant and growing ecosystem of public NWB data (**Tab. 1**) and tools utilizing NWB. The core NWB software stack has grown substantially since the inception of NWB 2.0 and has continued to grow steadily since the release of NWB 2.0 in January 2019, illustrating

the need for continued development and maintenance of NWB **(Fig. 2d).** In 2020 more than 600 scientists participated in NWB developer and user workshops and we have seen steady growth in attendance at NWB events over time (**Fig. 2e)**. NWB also provides extensive online training resources, including video and code tutorials, API and schema documentation, as well as guidelines and best practices (see **Supplement Material 2,3** and Methods). Community liaisons provide expert consultation for labs adopting NWB, CatalystNeuro can provide customized data conversion software for individual labs. As **Table 1** shows, despite its young age relative to the neurophysiology community, NWB 2.0 is being adopted by a growing number of diverse neuroscience laboratories and projects, creating a vibrant community where users can exchange and reuse data, with NWB as a common data standard[18].

Name, Affiliation	Species	Modality	DANDI datasets
Allen Institute	mouse, human	ecephys, icephys, ophys	12, 20, 21, 22, 23, 24, 30, 36, 37, 39, 42, 43, 48, 49, 50
R. Axel, Columbia	fly	ophys	
Blue Brain Project	mouse	icephys	25
J. Berke, UCSF	rat	ecephys	
K. Bouchard, LBNL/UC Berkeley	rat, simulation	ecephys, uECoG	
B. Brunton, U Washington	human	ECoG	55
E. Buffalo, U Washington	monkey	ecephys	
T. Buschman, Princeton	monkey	ecephys	
G. Buzsaki, NYU	rat, mouse	ecephys	3, 41, 44, 56, 59
M. Capogna, Aarhus	mouse	ecephys	
M. Carandini, UCL	mouse	ecephys	17
E. Chang, UCSF	human	ECoG	19
A. Churchland, CSHL	mouse	ecephys, ophys	16
D. Feldman, UC Berkeley	mouse	ecephys	
L. Frank, UCSF	mouse	ecephys	
L. Giocomo, Stanford	mouse	ecephys, ophys	53, 54
A. Groh, Heidelberg	mouse	ecephys, ophys	
K. Harris, UCL	mouse	ecephys	17

M. Hennig, Edinburgh	mouse	ecephys	28, 34
International Brain Lab	mouse	ecephys	45
D. Jaeger, Emory	mouse	ophys, ecephys, icephys	
S. Kastner, Princeton	monkey	ecephys	
N. Li, Baylor	mouse	ecephys	7
A. Losonczy, Columbia	mouse	ophys	
J. Martinez, Western	mouse, monkey, human	icephys	
D. O'Connor, Johns Hopkins	mouse	ecephys, ophys	
J. Parvizi, Stanford	human	ECoG	
U. Rutishauser, Cedars-Sinai	human	ecephys	4
B. Sabatini, Harvard	mouse	icephys	
S. Schultz, Imperial	mouse	ecephys, ophys	
I. Soltesz, Stanford	mouse, simulation	ophys, icephys	
N. Steinmetz, U Washington	mouse	ecephys	17
K. Svoboda, Janelia	mouse	ecephys, ophys	5, 6, 9, 10, 11, 13, 15
N. Tandon, UT Houston	human	ECoG	
D. Tank, Princeton	mouse	ecephys	
A. Tolias, Baylor	mouse	icephys	8, 35
S. Tripathy, UofToronto/CAMH	human, mouse	icephys	
T. Valiante, Toronto	human	icephys	

Table 1. Diverse Community of Data Producers Adopting NWB.

ecephys: extracellular electrophysiology; icephys: intracellular electrophysiology; ophys: optical physiology. All DANDI datasets can be found at https://dandiarchive.org/dandiset/{id}

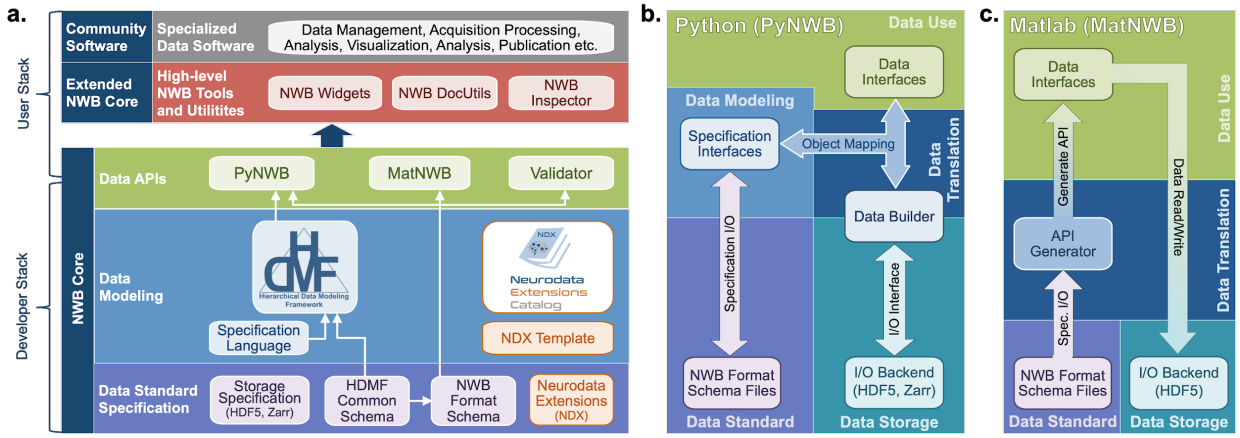


Figure 3: The NWB software architecture modularizes and integrates all components of a data language.

- **a.** Software stack for defining and extending the NWB data standard and creating and using NWB data files. The software stack covers all aspects of data standardization: i) data specification, ii) data modeling, iii) data storage, iv) data APIs, v) data translation, and vi) tools. Depending on their role, different stakeholders typically interact with different subsets of the software ecosystem. End users typically interact with the data APIs (green) and higher level tools (red, gray) while tool developers typically interact with the data APIs and data modeling layers (green, blue). Working groups and developers of extensions then typically interact with the data modeling and data standard specification components. Finally, core NWB developers typically interact with the entire developer stack, from foundational documents (lilac) to data APIs (green).
- **b.** Software architecture of the PyNWB Python API. PyNWB provides interfaces for interacting with the specification language and schema, data builders, storage backends, and data interfaces. Additional software components (arrows) insulate and formalize the transitions between the various components. The object mapping describes: 1) the integration of data interfaces (which describe the data) with the specification (which describes the data model) to generate data builders (which describe the data for storage) and 2) vice versa, the integration of data builders with the specification to create data interfaces. The object mapping insulates the end-users from specifics of the standard specification, builders, and storage, hence, providing stable, easy-to-use interfaces for data use that are agnostic of the data storage and schema. The I/O interface then provides an abstract interface for translating data builders to storage which specific I/O backends must implement. Finally, the specification I/O then describes the translation of schema files to/from storage, insulating the specification interfaces from schema storage details.
- **c.** Software architecture of the MatNWB Matlab API. MatNWB generates front-end data interfaces for all NWB types directly from the NWB format schema. This allows MatNWB to easily support updates and extensions to the schema while enabling development of higher-level convenience functions.

The NWB software architecture modularizes and integrates all components of a data language

The diversity of neurophysiology experiments presents an outstanding challenge to developing a language to communicate the experimental metadata. Furthermore, such a language must be instantiated by software that is usable, sustainable, and extensible. To achieve these software requirements, NWB defines and modularizes the interdependent,

yet separable, components of a data language. Defining and modularizing separable components is critical from a technological perspective to provide the flexibility required to address varying components independently. Additionally, the modularization addresses sociological challenges by allowing working groups, developer teams, and stakeholders to focus discussions and avoid conflicts due to overlap of component functionality.

The NWB data language is composed of five main components (**Fig. 3a**). First, data modeling seeks to describe scientific data standards and includes the NWB specification language for formal description of hierarchical data standards. NWB uses the Hierarchical Data Modeling Framework (HDMF)[17] for creating hierarchical data standards and Application Programming Interfaces (APIs)**,** and the Neurodata Extension Catalog and associated tools for managing extensions to the NWB data standard. Next, a data specification describes the organization of complex scientific data and includes the NWB standard schema, governing documents, and extensions. To organize complex neurophysiology data, the NWB schema builds on common data structures defined by the HDMF common schema. Next, the data storage component deals with translating NWB data models to/from storage on disk. Data storage is governed by formal specifications describing the translation of NWB data primitives to primitives of the particular storage backend format (e.g., HDF5) and is implemented as part of the NWB user APIs. The data use component concerns methods and interfaces to enable effective interactions with NWB data (e.g., for read, write, query, and validation) and are instantiated by the PyNWB and MatNWB reference APIs. Finally, data translation is provided by the NWB API's (**Fig. 3b,c**) and include methods that insulate and integrate the various components by providing structured interfaces for translating between user data interfaces, storage, and specifications.

The PyNWB and MatNWB APIs provide interfaces to NWB for the user implemented in the two most common data science programming languages in neuroscience (Python and Matlab respectively). These APIs (**Fig. 3b,c**) are governed by the NWB schema and use an object-oriented design in which neurodata types (e.g., ElectricalSeries, OpticalSeries) are represented by a dedicated interface class. Both APIs are interoperable (i.e., files generated by PyNWB can be read via MatNWB and vice

versa) and support advanced data Input/Output (I/O) features, such as lazy data read, compression, and iterative data write for data streaming. A key difference in the design of PyNWB and MatNWB is the implementation of the data translation process. PyNWB builds on HDMF[17], which implements a dynamic data translation process based on data builders (**Fig. 3b**). The data builders define abstract descriptions of basic primitives of the NWB specification language (e.g., Groups, Datasets, Attributes, Links, etc.,). The object mapping then uses the format schema to translate user data defined in the data interfaces to data builders for storage and conversely to instantiate data interfaces based on data read from storage. PyNWB provides custom interfaces for all NWB neurodata types, and in cases where no custom data interface classes are available (e.g., in the case a user uses an extension), the API supports automatic, dynamic generation of interfaces classes at runtime. This dynamic data translation approach further insulates the data storage, data specification, and data interface components. This approach facilitates the integration of new technologies (e.g., alternate data storage backends based on Zarr or databases) with NWB. In contrast, MatNWB implements a static translation process that uses API generators to automatically generate Matlab classes (**Fig. 3c)**. This approach simplifies updating of the API to support new versions of the NWB schema and extensions, and helps minimize cost for development. The difference in the data translation process between the APIs (i.e., static vs. dynamic) is a reflection of the different target uses (**Fig. 3b,c**). While MatNWB primarily targets data conversion and analysis, PyNWB also targets integration with data archives and web technologies and is used heavily for development of extensions and exploration of new technologies, such as alternate storage mechanisms and parallel computing libraries.

Building on the core NWB software, a growing set of high-level tools are available as part of the extended NWB core, dedicated to providing additional functionality and to simplify interaction with NWB (**Fig. 3a**). The extended core includes NWB-specific tools, e.g., NWB Widgets for visualization and exploration of NWB files or the NWB DocUtils for documenting extensions. All NWB software is managed and versioned using Git and released using a permissive BSD license via GitHub. NWB uses automated continuous integration for testing on all major architectures (MacOS, Windows, and Linux) and all core software can be installed via common package managers (e.g., pip and conda). The

suite of NWB core software tools enable users to easily read and write NWB files, extend NWB to integrate new data types, and build the foundation for integration of NWB with community software. NWB data can also be easily accessed in other programming languages (e.g., IGOR or R) using the HDF5 APIs available across modern scientific programming languages.

With its flexible design and open model, the NWB software stack instantiates the NWB data language and makes NWB accessible to users and developers. The NWB software model both illustrates the complexity of creating a data language and provides reusable components, e.g., HDMF, that can be applied more broadly to facilitate development of data languages for other biological fields in the future.

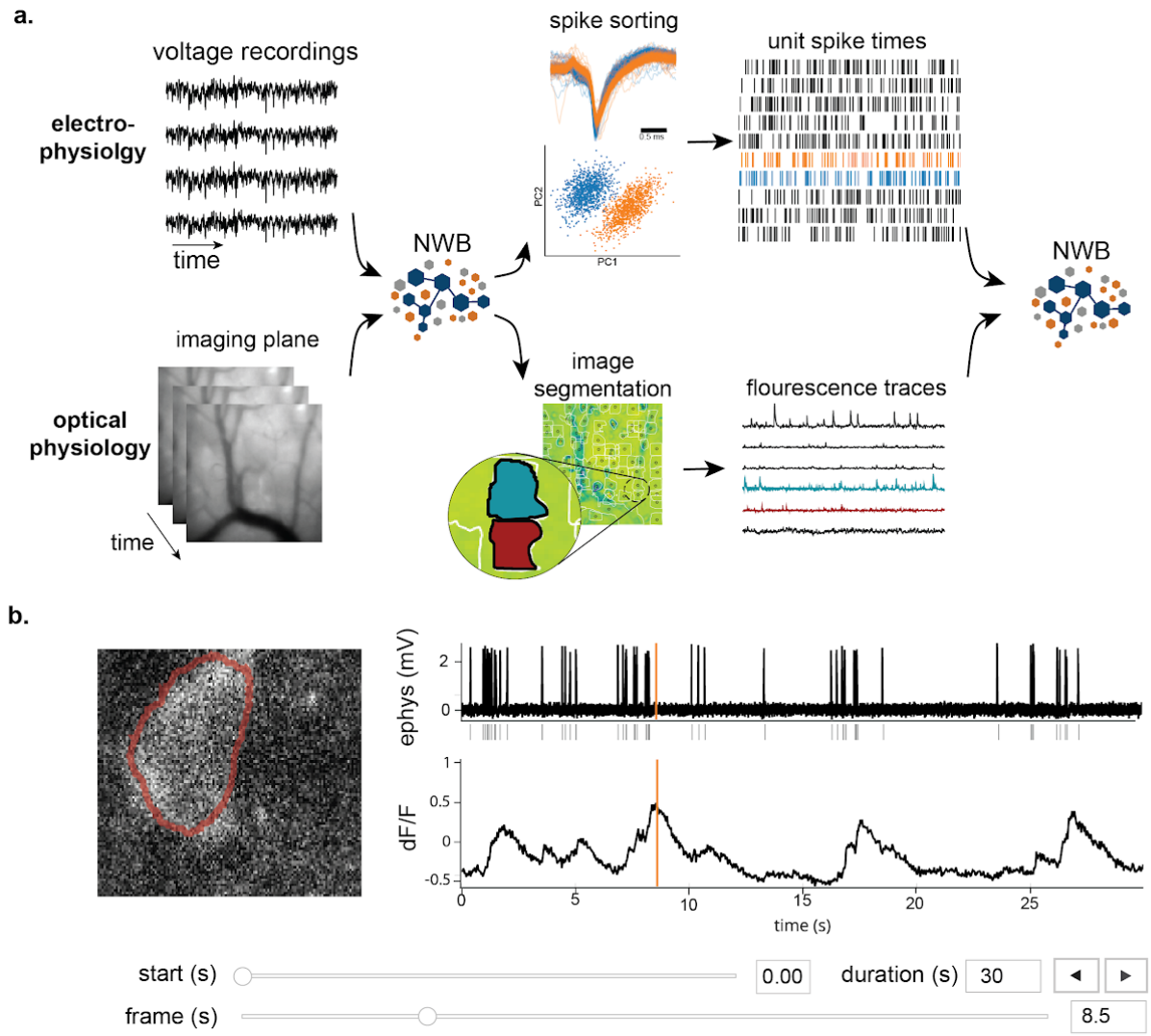


Figure 4. NWB enables unified description and storage of multimodal raw and processed data

a. Example pipelines for extracellular electrophysiology and optical physiology demonstrate how NWB facilitates data processing. For extracellular electrophysiology (top), raw acquired data is written to the NWB file. The NWB ecosystem provides interfaces to a variety of spike sorters that extract unit spike times from the raw electrophysiology data. The spike sorting results are then stored into the same NWB file (bottom). Separate experimental data acquired from an optical technique is converted and written to the NWB file. Several modern software tools can then be used to process and segment this data, identifying regions that correspond to individual neurons, and outputting the fluorescence trace of each putative neuron. The fluorescence traces are written to the same NWB file. NWB handles the time alignment of multiple modalities, and can store multiple modalities simultaneously, as shown here. The NWB file also contains essential metadata about the experimental preparation.

b. NWBWidgets provides visualizations for the data within NWB files with interactive views of the data across temporally aligned data types. Here, we show an example dashboard for simultaneously recorded electrophysiology and imaging data. This interactive dashboard shows on the left the acquired image and the outline of a segmented neuron (red) and on the right a juxtaposition of extracellular electrophysiology, extracted spike times, and simultaneous fluorescence for the

segmented region. The orange line on the ephys and dF/F plots indicate the frame that is shown to the left. The controls shown at the bottom allow a user to change the window of view and the frame of reference within that window.

NWB enables unified description and storage of multimodal data and derived products.

Neurophysiology experiments often contain multiple simultaneous streams of data. For instance, a single experiment might contain simultaneous neural recordings, sensory stimuli, behavioral tracking, and direct neural activation or inhibition. Furthermore, scientists are increasingly leveraging multiple neurophysiology recording modalities simultaneously (e.g., ephys and ophys), which offer complementary information not achievable in a single modality. Each of these distinct data input types require custom processing at multiple stages, further expanding the multiplicity of diverse data types that need to be represented. Current formats in neuroscience focus on a single modality, leaving it to the scientist to harmonize diverse data streams, which is laborious and error-prone.

A key capability of NWB is to describe and store many data sources in a unified format that is ready to analyze with all time bases aligned to a common clock. These data sources include neurophysiological recordings, behavior, and stimulation information. For each of these modalities, raw acquired signals and/or preprocessed data can be stored in the same file. **Figure 4a** illustrates a dataset where electrophysiology and optical physiology were recorded simultaneously, and stored in NWB[19,20]. The raw voltage traces (**Fig. 4a, top**) from the electrophysiology recordings and image sequences from the optical recordings (**Fig. 4a, bottom**) can both be stored in the same NWB file, or separate NWB files synchronized to each other. Electrophysiology data generally goes through a spike sorting process which processes the voltage traces into individual units and action potential times for those units (**Fig. 4a, top**). These unit spike times can then also be written to the NWB file. Similarly, optical physiology generally is processed using image segmentation to identify regions of the image that correspond to units and extract fluorescence traces for each unit (**Fig. 4a, bottom**). These fluorescence traces can also be stored in the NWB file, resulting in raw and processed data for multiple input streams. As illustrated in **Figure 1d**, NWB can also store raw and processed behavioral data as well as stimuli, such as animal location and amplitude/frequency of sounds. The multi-

modal capability of NWB is critical for capturing the diverse types of data simultaneously acquired in standard neurophysiology experiments, particularly if those experiments involve multiple simultaneous neural recording modalities.

Pre-synchronized streams enable faster and less error-prone development of visualizations, such as simultaneous views across multiple streams. **Figure 4b** shows an interactive dashboard for exploring a separate dataset of simultaneous recorded optical physiology and electrophysiology data published by the Allen Institute[19]. This dashboard illustrates the simultaneous exploration of five data elements all stored in a single NWB file. The microscopic image panel (**Fig. 4b, far left)** shows a frame of the video recorded by the microscope. The red outline overlaid on that image shows the region-of-interest where a cell has been identified by the experimenter. The fluorescence trace (dF/F) shows the activation of the region-of-interest over time. This activity is displayed in line with electrophysiology recordings of the same cell (ephys), and extracted spikes (below ephys). Interactive controls **(Fig. 4b, bottom)** allow a user to explore the complex and important relationship between these data sources.

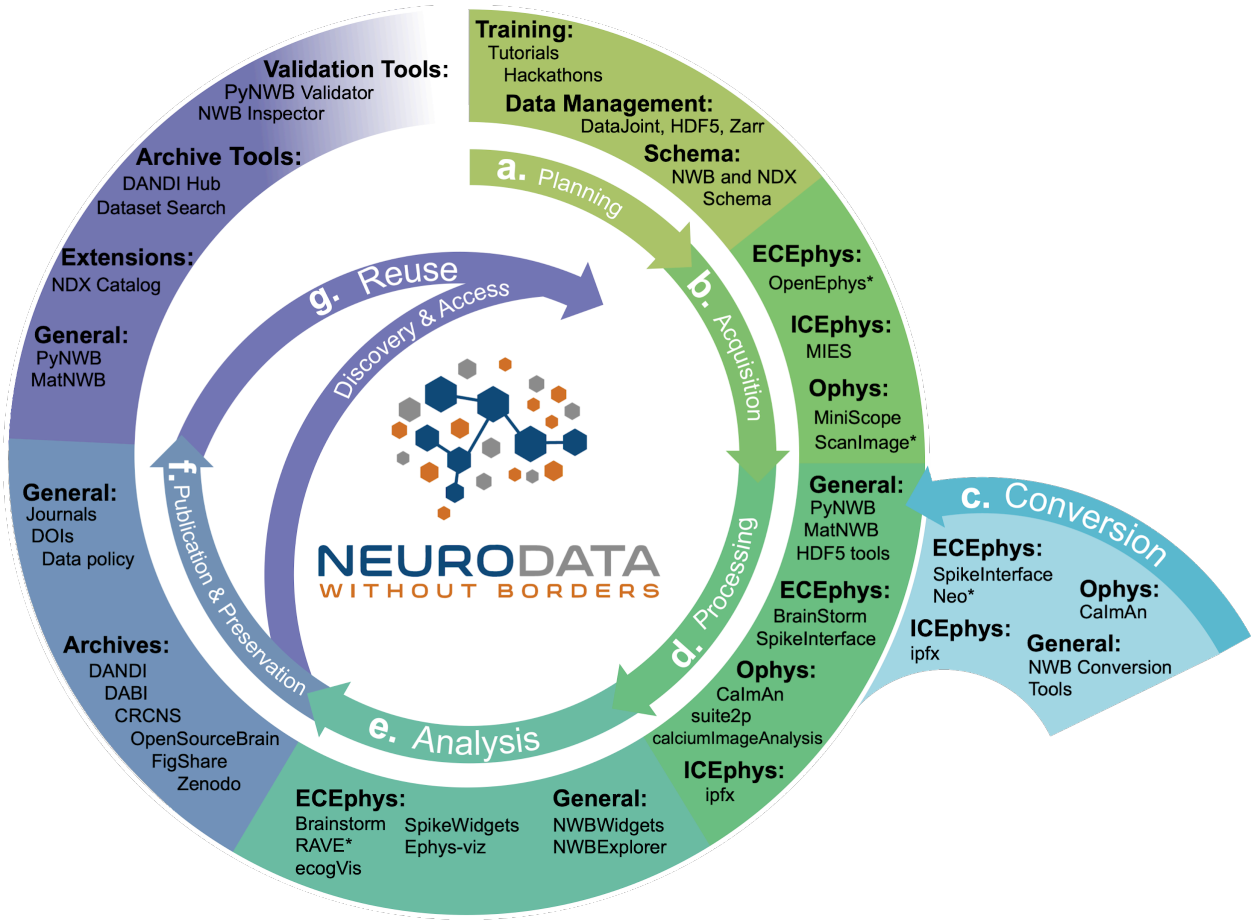


Figure 5: NWB is integrated with state-of-the-art analysis tools throughout the data life-cycle NWB technologies are at the heart of the neurodata lifecycle and applications. Data standards are a critical conduit that facilitate the flow of data throughout the data lifecycle and integration of data and software across all phases (a. to g.) of the data lifecycle.

- **a.** NWB supports experimental planning through integration with data management, best practices, and by allowing users to clearly define what metadata to collect.
- **b. c.** NWB supports storage of unprocessed acquired electrical and optical physiology signals, facilitating integration already during data acquisition. NWB is already supported by several acquisition systems (a) as well as a growing set of tools for conversion (d) of existing data to NWB.
- **d.** Despite its young age, NWB is already supported by a large set of neurophysiology processing software and tools. Being able to access and evaluate multiple processing methods, e.g, different spike sorting algorithms and ROI segmentation methods, is important to enable high-quality data analysis. Through integration with multiple different tools, NWB provides access to broad range of spike sorters, including, MountainSort, KiloSort, WaveClust, and others, and ophys segmentation methods, e.g, CELLMax, CNMF, CNMF-E, and EXTRACT.
- **e.** For scientific analysis, numerous general tools for exploration and visualization of NWB files (e.g,. NWBWidgets and NWBExplorer) as well as application-specific tools for advanced analytics (e.g., Brainstorm) are accessible to the NWB community.
- **f. g.** NWB is supported by a growing set of data archives for publication and preservation of research data. Data archives in conjunction with NWB APIs, validation tools, and the NDX Catalog also play a central role in facilitating data reuse and discovery.

NWB integrates with (not competes with) existing and emerging technologies across the data lifecycle, creating a flourishing NWB software ecosystem that enables users to access state-of-the-art analysis tools, share and reuse data, improve efficiency and reduce cost, and accelerate and enable the scientific discovery process.

NWB is integrated with state-of-the-art analysis tools throughout the data life-cycle

Scientific data should not be considered in isolation, but must be thought of in the context of the entire data life-cycle. The data life-cycle spans planning, acquisition, processing, and analysis to publication and reuse (**Fig. 5**). Through its rigorous schema, best software engineering practices and implementation, and user training, NWB provides a common language for neurophysiology data collected using existing and emerging neurophysiology technologies integrated into a vibrant neurophysiology data ecosystem. This is critical to truly make neurophysiology data Findable, Accessible, Interoperable, and Reusable (i.e., FAIR)[21].

NWB supports experimental planning by helping users to clearly define what metadata to collect and how the data will be formatted and managed (**Fig. 5a**). To support data acquisition, NWB allows for the storage of unprocessed acquired electrical and optical physiology signals. Storage of these signals requires either streaming the data directly to the NWB file from the acquisition system or converting data from other formats after acquisition (**Fig. 5b,c**). Some acquisition systems, such as the MIES[22] intracellular electrophysiology acquisition platform, already support direct recording to NWB and the community is actively working to expand support for direct recording to NWB, e.g., via ScanImage[23] and OpenEphys[24]. To support integration of legacy data and other acquisition systems, a variety of tools exist for converting neurophysiology data to NWB. For extracellular electrophysiology, the SpikeInterface package[25] provides a uniform API for conversion and processing data that supports conversion for 19+ different proprietary acquisition formats to NWB. For intracellular electrophysiology, the Intrinsic Physiology Feature Extractor (IPFX) package[26] supports conversion of data acquired with Patchmaster. Direct conversion of raw data to NWB at the beginning of the data lifecycle facilitates data re-processing and re-analysis with up-to-date methods and data re-use more broadly.

The NWB community has been able to grow a flourishing ecosystem of software tools that offer convenient methods for processing data from NWB files (and other formats) and writing the results into an NWB file (**Fig. 5d**). NWB allows these tools to be easily accessed and compared, and used interoperably. Furthermore, storage of processed data in NWB files allows direct re-analysis of activation traces or spike times

via novel analysis methods without having to reproduce time-intensive pre-processing steps. For example, the SpikeInterface API supports export of spike sorting results to NWB across nine different spike sorters and customizable data curation functions for integration of results from multiple spike sorters with common metrics. For optical physiology, several popular state-of-the-art software packages, such as CaImAn[27], suite2p[28], and ciapkg[29], help users build processing pipelines that segment optical images into regions of interest corresponding to putative neurons, and write these results to NWB.

We also see a range of general and application-specific tools emerging for analysis of neurophysiology data in NWB (**Fig. 5e**). The NWBWidgets[30] library enables interactive exploration of NWB files via interactive views of the NWB file hierarchy and dynamic plots of neural data, e.g., visualizations of spike trains and optical responses. NWB Explorer[31] developed by MetaCell in collaboration with OpenSourceBrain is a web app that allows a user to explore any publicly hosted NWB file and supports custom visualizations, analysis via Jupyter notebooks, as well as use of the NWBWidgets. These tools allow scientists to inspect their own data for quality control, and allow data reusers to quickly understand the contents of a published NWB file. In addition to these general-purpose tools, many application-specific tools, e.g., RAVE[32], ecogVIS[33], Brainstorm[34], Neo[35] and others are already supporting or are in the process of developing support for analysis of NWB files.

Many journals and funding agencies are beginning to require that data be made FAIR. For publication and preservation (**Fig. 5f**) archives are an essential component of the NWB ecosystem, allowing data producers to document data associated with publications and share that data with others. NWB files can be stored in many popular archives, such as FigShare and Collaborative Research in Computational Neuroscience (CRCNS.org). In the context of the NIH BRAIN Initiative, the DANDI archive has been specifically designed to publish and validate NWB files and leverage their structure for searching across datasets. In addition, several other archives, e.g. DABI[36] and OpenSourceBrain[37], are also supporting publication of NWB data.

Data archives also play a crucial role in discovery and reuse of data (**Fig. 5g**). In addition to providing core functionality for data storage and search, archives increasingly also provide compute capacity for reanalysis. For example, with DANDI Hub, the DANDI archive provides users a familiar Jupyter Hub interface that supports interactive

exploration and processing of NWB files stored in the archive. The NWB data APIs, validation, and inspection tools also play a critical role in data reuse to enable access and assure data validity. Finally, the Neurodata Extension Catalog described below facilitates accessibility and reuse of data files that use NWB extensions.

Thus, NWB provides a common language to describe neurophysiology experiments, data, and derived data products (e.g., outputs of spike-sorting) that enables users to maintain and exchange data throughout the data lifecycle and access state-of-the-art software tools. Together the tools in the NWB ecosystem make neurophysiology data much more FAIR.

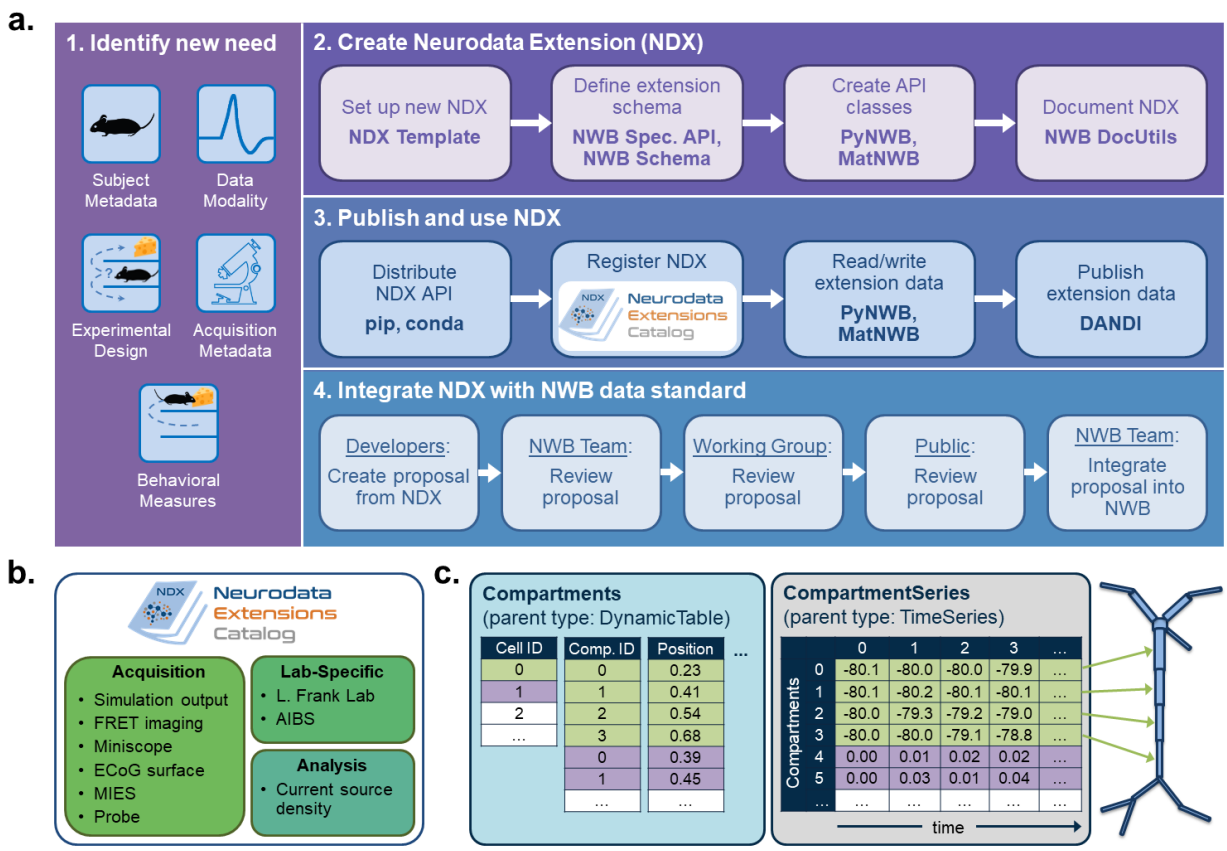


Figure 6: NWB enables creation and sharing of extensions to incorporate new use cases

a. Schematic of the process of creating a new neurodata extension (NDX), sharing it, and integrating it with the core NWB data standard. Users first identify the need for a new data type, such as additional subject metadata or data from a new data modality. Users can then use the NDX Template, NWB Specification API, PyNWB/MatNWB data APIs, and NWB DocUtils tools to set up a new NDX, define the extension schema, define and test custom API classes for interacting with

extension data, and generate Sphinx-based documentation in common formats, e.g., HTML, PDF. After the NDX is completed, users can publish the NDX on PyPI and conda-forge for distribution via the pip and conda tools, and share extensions via the NDX Catalog, a central, searchable catalog. Users can easily read/write extension data using PyNWB/MatNWB and publish extension data in DANDI and other archives. Finally, extensions are used to facilitate enhancement, maintenance, and governance of the NWB data standard. Users may propose the integration of an extension published in the NDX Catalog with the core standard. The proposal undergoes three phases of review: an initial review by the NWB technology team, an evaluation by a dedicated working group, and an open, public review by the broader community. Once approved, the proposal is integrated with NWB and included in an upcoming version release.

b. Sampling of extensions currently registered in the NDX catalog. Users can search extensions based on keywords and textual descriptions of extensions. The catalog manages basic metadata about extensions, enabling users to discover and access extensions, comment and make suggestions, contribute to the source code, and collaborate on a proposal for integration into the core standard. While some extensions have broad applicability, others represent data and metadata for a specific lab or experiment.

c. Example extension for storing simulation output data using the SONATA framework. The new Compartments type extends the base DynamicTable type and contains metadata about each cell and compartment within each cell, such as position and label. The CompartmentSeries type extends the base TimeSeries type and contains a link to the Compartments type to associate each row of its data array with a compartment from the Compartments table.

NWB enables creation and sharing of extensions to incorporate new use cases

As with all of biology, neurophysiological discovery is driven in large part by individual scientists conducting novel experiments with new tools to answer previously unconsidered questions. Furthermore, neurophysiology experiments are already highly diverse (e.g., **Fig. 1a-b**) and can require the storage of new data types and additional metadata. For example, an individual lab may need to store metadata about the subject's environment or training history, data from lab-specific calibration tests, or metadata specific to particular devices. Thus, a language for neurophysiology data must be able to co-evolve with the experiments being performed and provide customization options to the degree possible.

NWB enables the creation and sharing of user-defined extensions to the standard that support new and specialized data types (**Fig. 6a1**). Neurodata extensions (NDX) are defined using the same formal specification language used by the core NWB schema. Importantly, extensions can build off of data types defined in the core schema or other extensions through inheritance and composition. This enables the reuse of definitions and associated code, facilitates the integration with existing tools, and makes it easier to contextualize new data types.

NWB provides a comprehensive set of tools and services for developing and using neurodata extensions. NWB tools, such as the NWB Specification API, NWB DocUtils,

and the PyNWB and MatNWB user APIs work with extensions with little adjustment (**Fig. 6a2**). In addition, the NDX Template makes it easy for users to develop new extensions, and the Neurodata Extensions Catalog (**Fig. 6a3**) provides a centralized listing of extensions for users to publish, share, find, and discuss extensions across the community. Several extensions have been registered in the Neurodata Extensions Catalog (**Fig. 6b**), including extensions to support the storage of the cortical surface mesh of an electrocorticography experiment subject, storage of fluorescence resonance energy transfer (FRET) microscopy data, and metadata from an intracellular electrophysiology data acquisition system. The catalog also includes the ndx-simulation-output extension for the storage of the outputs of large-scale simulations. Large-scale network models that are described using the new SONATA format[38] can be converted to NWB using this extension.

As particular extensions gain traction within the community, they may be integrated into the core NWB format for broader use and standardization (**Fig. 6a4**). NWB has a formal, community-driven review process for refining the core format so that NWB can adapt to evolving data needs in neuroscience. The owners of the extension can submit a community proposal for the extension to the NWB Technical Advisory Board, which then evaluates the extension against a set of metrics and best practices published on the catalog website. The extension is then tested and reviewed by both a dedicated working group of potential stakeholders and the general public before it is approved and integrated into the core NWB format. Key advantages of the extension approach are that it allows iterative development of extensions and complete implementation and vetting of new data types under several use cases before they become part of the core NWB format. The NWB extension mechanism thus enables NWB to provide a unified data language for all data related to an experiment, allows sharing of data from novel experiments, and supports the process of evolving the core NWB standard to fit the needs of the neuroscientific community.

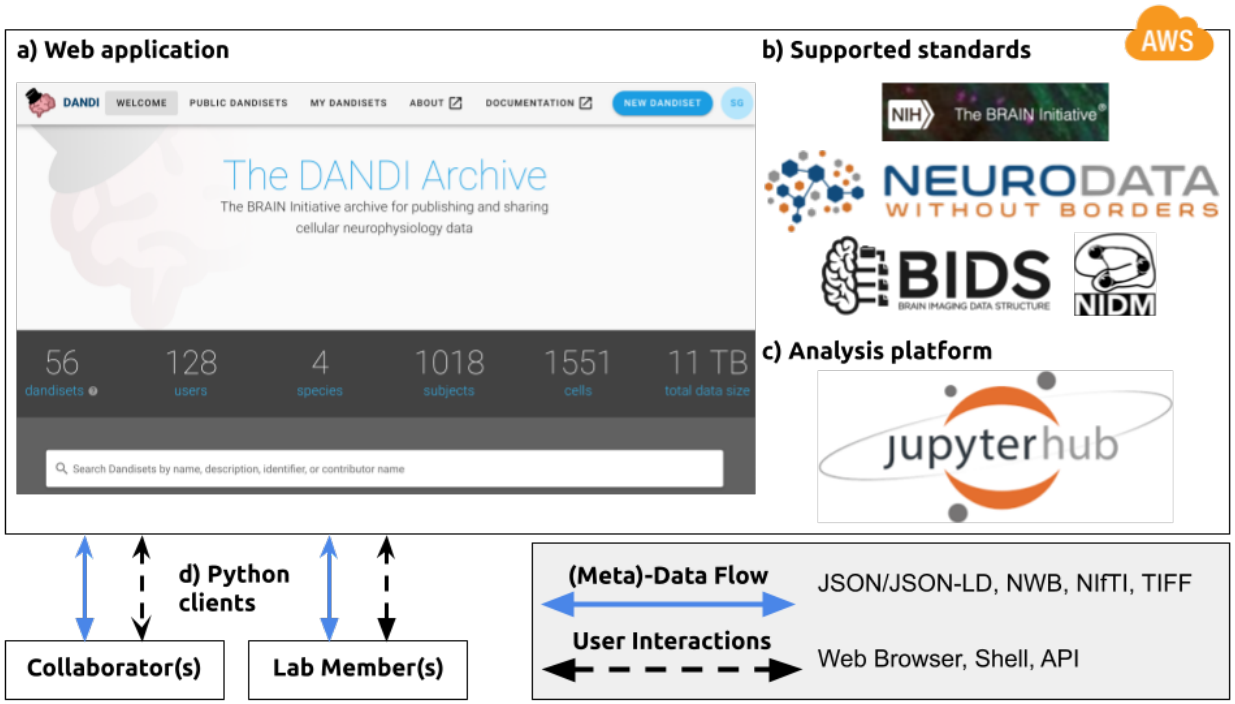


Figure 7. NWB is foundational for the DANDI data repository to enable collaborative data sharing.
The DANDI project makes data and software for cellular neurophysiology FAIR. DANDI stores electrical and optical cellular neurophysiology recordings and associated MRI and/or optical imaging data. NWB is foundational for the DANDI data repository to enable collaborative data sharing.

- **a.** DANDI provides a Web application allowing scientists to share, collaborate, and process data from cellular neurophysiology experiments. The dashboard provides a summary of Dandisets, and allows users to view details of each dataset.
- **b.** DANDI works with US BRAIN Initiative awardees and the neurophysiology community to curate data using community data standards such as NWB, BIDS, and NIDM. DANDI is supported by the US BRAIN Initiative and the Amazon Web Services (AWS) Public Dataset Program.
- **c.** DANDI provides a Jupyterhub interface to visualize the data and interact with the archive directly through a browser, without the need to download any data locally.
- **d.** Using Python clients and/or a Web browser researcher can submit and retrieve standardized data and metadata from the archive. The data and metadata use standard formats such as JSON, JSON-LD, NWB, NIfTI, and TIFF.

NWB is foundational for the DANDI data repository to enable collaborative data sharing

With the expansion of neurophysiology hardware, extended recording options, and increasing complexity of experiments, neuroscientists are generating data at unprecedented scale. Making such data accessible supports published findings and allows secondary reuse. To date, many neurophysiology datasets have been deposited into a diverse set of repositories (e.g., CRCNS, Figshare, Open Science Framework, Gin).

However, no single data archive provides to the neuroscientific community the capacity and the domain specificity to store and access large neurophysiology datasets. Most current repositories have specific limits on data sizes, are often generic, and therefore lack the ability to search using domain specific metadata. Further, for most neuroscientists, these archives often serve as endpoints associated with publishing, while research is typically an ongoing and collaborative process. Few data archives today support a collaborative research model that allows data submission prospectively, and opening the conversation to a broader community. For the US BRAIN Initiative, this is also a mandate. Together, these issues impede access and reuse of data, ultimately decreasing the return on investment into data collection by both the experimentalist and the funding agencies.

To address these and other challenges associated with neurophysiology data storage and access, we developed DANDI, a Web-based data archive that also serves as a collaboration space for neurophysiology projects (**Fig. 7**). The DANDI data archive (https://dandiarchive.org) is a cloud-based repository for cellular neurophysiology data and uses NWB as its core data language (**Fig. 7b**). Users can organize collections of NWB files (e.g., recorded from multiple sessions) into DANDI datasets (so called Dandisets). Users can view the public Dandisets using a Web browser (**Fig. 7a**) and search for data from different projects, people, species and modalities. They can interact with the data in the archive using a Jupyterhub Web interface (**Fig. 7c**). Using the DANDI Python client, users can organize data locally into the structure required by DANDI as well as download data from and upload data to the archive (**Fig. 7d**). Software developers can access information about Dandisets and all the files it contains using the DANDI Representational State Transfer (REST) API (https://api.dandiarchive.org/). The REST API also allows developers to create software tools and database systems that interact with the archive. Each Dandiset is structured by grouping files belonging to different biosamples, with some relevant metadata stored in the name of each file, and thus aligning itself with the BIDS standard[39]. Metadata in DANDI is stored using the JSON-LD format, thus allowing graph-based queries and exposing DANDI to Google Dataset Search. Dandiset creators can use DANDI as a living repository and continue to add data and analyses to an existing Dandiset. Released versions of Dandisets are immutable and

receive a digital object identifier (DOI). The data are presently stored on an Amazon Web Services Public dataset program bucket, enabling open access to the data over the Web, and backed up on institutional repositories. DANDI is working with hardware platforms (e.g., OpenEphys), database software (e.g., DataJoint) and various data producers to generate and distribute NWB datasets to the scientific community. The current data statistics can be found at https://dandiarchive.org.

DANDI provides neuroscientists and software developers with a Platform as a Service (PAAS) infrastructure based on the NWB data language and supports interaction via the web browser or through programmatic clients, software, and other services. In addition to serving as a data archive and providing persistence to data, it supports continued evolution of Dandisets. This allows scientists to use the archive to collect data towards common projects across sites, and engage collaborators actively, directly at the onset of data collection rather than at the end. DANDI also provides a computational interface to the archive to accelerate analytics on data, and links these Dandisets to eventual publications when generated. The code repositories for the entire infrastructure are available on Github[40] under an Apache 2.0 license.

Discussion

To investigate the myriad functions of the brain across species necessitates a massive diversity and complexity of neurophysiology experiments. This diversity presents an outstanding barrier in meaningful sharing and collaborative analysis of the collected data, and ultimately prevents the data from being FAIR. To overcome this barrier, we developed a data ecosystem based on the Neurodata Without Borders (NWB) data language and software. NWB is being utilized by more than 35 labs to enable unified storage and description of intracellular, extracellular, LFP, ECoG, and Ca^{2+} data in fly, mice, rats, monkeys, humans, and simulations. To support the entire data lifecycle, NWB natively operates with processing, analysis, visualization, and data management tools, as exemplified by the ability to store both raw and pre-processed simultaneous electrophysiology and optophysiology data. Formal extension mechanisms enable NWB to co-evolve with the needs of the community. Finally, NWB enables DANDI to provide a data archive that also serves as a collaboration space for neurophysiology projects.

NWB as the lingua franca of neurophysiology data

The diversity of experiments, data types, and the rate of methodological innovation are key impediments to adoption of data standards in scientific research practice. Addressing this challenge requires a paradigm shift in how we conceptualize the solution. Scientists need more than a rigid data format, but instead require a flexible data language. Such a language should enable scientists to communicate via data. Natural languages evolve with the concepts of the societies that use them, while still providing a basis that enables communication of common concepts. Similarly, a scientific data language should evolve with the scientific research community, and at the same time provide a standardized core that expresses common and established methods and data types. NWB is such a data language for neurophysiology experiments.

The NWB specification language provides, much like an alphabet and core grammatical rules in natural language, the basic tools and rules that allow us to create words and phrases. The format schema, much like a dictionary and grammatical rules for sentence and document structure, provides us with the words and phrases (neurodata_types) of our data language and rules for how to compose them to form data documents (NWB files). And much like we store natural language in many different mediums (e.g., via printed books, electronic records, or handwritten notes), flexible data storage methods allow us to manage and share data in different forms depending on the application. User APIs (here HDMF, PyNWB, and MatNWB) provide us, much like text editors for natural language, with tools to create, read, and modify data documents and interact with core aspects of the language. NWB Extensions provide a mechanism to create, publish, and eventually integrate new modules into NWB to ensure it co-evolves with the tools and needs of the neurophysiology community. Finally, much like a bookstore or wikipedia, DANDI provides a cloud-based platform for archiving, sharing, and collaborative analysis of NWB data. Together, these interacting components provide a comprehensive data ecosystem for the neuroscience community that enables reproduction, interchange, and reuse of diverse neurophysiology data.

NWB is community driven, professionally developed, and democratizes neurophysiology

Today, there is an enormous diversity in data formats and non-interoperable tools used by the neuroscience community. Often, these formats and tools are specific to the lab and even the researcher. This level of specificity is a major impediment to sharing data and reproducing results, even within the same lab. More broadly, this fragmentation of the data space reinforces siloed research, and makes it difficult for datasets or software to be impactful on a community level. Our goal is for the NWB data language to be foundational in deepening collaborations between the community of neuroscientists. The current NWB software is the result of an intense, community lead, years-long collaboration between experimental and computational neuroscientists with data scientists and computer scientists. NWB is governed by a similarly diverse group to ensure both the integrity of the software and that NWB continues to meet the needs of the neuroscience community.

As with all sophisticated scientific instruments, there is some training required to get a lab's data into NWB. Several training and outreach activities provide diverse opportunities for the neuroscience community to utilize NWB. Tutorials, hackathons and user training events allow us to bring together neuroscientists who are passionate about open data and data management. These users bring their own data to convert or their own tools to integrate, which in turn makes the NWB community more diverse and representative of the overall neuroscience community. Our digital presence has allowed us to grow our community internationally and at an exponential rate. Updates on Twitter and our website (www.nwb.org), tutorials on Youtube, and free virtual hackathons, all are universally accessible and have helped us achieve a global reach, interacting with scientists from countries that are too often left out. Together, these outreach activities combined with NWB and DANDI democratizes both neurophysiology data and analysis tools, as well as the extracted insights.

The Future of NWB

To address the next frontier in grand challenges associated with understanding the brain, the neuroscience community must integrate information from diverse experiments spanning several orders of magnitude in spatial and temporal scales[13,14,41]. Currently, different domains of neuroscience (e.g., genomic/transcriptomic, anatomy,

neurophysiology, etc.,) are supported by standards that are not coordinated. Building bridges across neuroscience domains will necessitate interaction between the standards, and will require substantial future efforts. This will bring these communities, and the tools they develop, closer together. There are nascent activities for compatibility between NWB and the Brain Imaging Data Structure (BIDS), e.g., as part of the BIDS human intracranial neurophysiology ECoG/iEEG extension[42], but further efforts in this area are needed.

As adoption of NWB continues to grow, new needs and opportunities for further harmonization of metadata arise. For example, the NWB working group on ICEphys seeks to expand NWBs existing data structures to enable the description of the hierarchical organization of ICEphys experiments and associated metadata[43]**.** A key ongoing focus area is on development and integration of ontologies with NWB to enhance specificity, accuracy, and interpretability of data fields, e.g., the NWB working groups on genotype and spatial coordinate representation or the INCF Electrophysiology Stimulation Ontology working group[44]. Another key area is on extending NWB to new areas, such as, the ongoing working groups on integration of behavioral task descriptions with NWB (e.g., based on BEADL[45]) and enhanced integration of simulations with NWB.

Because of the diversity of neurophysiology experiments, it is notoriously challenging to make the resulting data FAIR. Together, the NWB data language and the NWB-based DANDI data archive, create a data science ecosystem for neurophysiology. This is a fantastic start to enhance the FAIRness of neurophysiology data. Underlying the cohesion of this ecosystem is a common language for the description of data and experiments: NWB. However, like all languages, NWB must continue to adapt to accommodate advances in neuroscience technologies and the evolving community using that language. We strongly advocate for funding support of all aspects of the data-software life-cycle (development, maintenance, integration, and distribution) to ensure the neuroscience community fully reaps the benefits of investment into neurophysiology tools and data acquisition.

Core design principles and technologies for biological data languages

The problems addressed by NWB technologies are not unique to neurophysiology data. Indeed, as was recently discussed in[46], lack of standards in genomics data is threatening

the promise of that data. Many of the tools and concepts of the NWB data language can be applied to enhance standardization and exchange of data in biology more broadly. For example, the specification language, HDMF, the concept of extensions and the extension catalog are general, broadly applicable technologies. The impact of the methods and concepts we have described here, hence, has the potential to extend well beyond the boundaries of neurophysiology. Indeed, funded projects are underway using HDMF to create a data standard for machine learning training results, HDMF-ML.

We developed design and implementation principles to create a robust, extensible, maintainable, and usable data ecosystem that embraces and enables the diversity of neurophysiology data. Across biology, experimental diversity and data heterogeneity are the rule, not the exception[2]. Indeed, as biology faces the daunting frontier of understanding life from atoms to organisms, the complexity of experiments and multimodality of data will only increase. Therefore, the principles developed and deployed by NWB may provide a blueprint for creating data ecosystems across other fields of biology.

Methods

NWB GitHub Organizations

All NWB software is available open source via the following three GitHub organizations. The Neurodata Without Borders[47] GitHub organization is used to manage all software resources related to core NWB software developed by the NWB developer community, e.g., the PyNWB and MatNWB reference APIs. The HDMF development[48] Github organization is used to publish all software related to the Hierarchical Data Modeling Framework (HDMF), including, HDMF, HDMF DocUtils, HDMF Common Schema and others. Finally, the NWB Extensions[49] GitHub organization is used to manage all software related to the NDX Catalog, including all extension registrations. Note, the catalog itself only stores metadata about NDXs, the source code of NDXs are often managed by the creators in dedicated repositories in their own organizations.

HDMF Software

HDMF software is available on GitHub using an open BSD licence model.

Hierarchical Data Modeling Framework (HDMF) is a Python package for working with hierarchical data. It provides APIs for specifying data models, reading and writing data to different storage backends, and representing data with Python objects. HDMF builds the foundation for the implementation of PyNWB and specification language. **[Source]**[50] **[Documentation]**[51] **[Web]**[52]

HDMF Documentation Utilities (hdmf-docutils) are a collection of utility tools for creating documentation for data standard schema defined using HDMF. The utilities support generation of reStructuredText (RST) documents directly from standard schema which can be compiled to a large variety of common document formats (e.g., HTML, PDF, epub, man and others) using Sphinx. **[Source]**[53]

HDMF Common Schema defines a collection of common reusable data structures that build the foundation for modeling of advanced data formats, e.g., NWB. APIs for the HDMF common data types are implemented as part of the `hdmf.common` module in the HDMF library. **[Source]**[54] **[Documentation]**[55]

HDMF Schema Language provides an easy-to-use language for defining hierarchical data standards. APIs for creating and interacting with HDMF schema are implemented in HDMF. **[Documentation]**[56]

NWB Software

NWB software is available on GitHub using an open BSD licence model.

PyNWB is the Python reference API for NWB and provides a high-level interface for efficiently working with Neurodata stored in the NWB format. PyNWB is used by users to create and interact with NWB and neuroscience tools to integrate with NWB. **[Source]**[57] **[Documentation]**[58]

MatNWB is the MATLAB ® reference API for NWB and provides an interface for efficiently working with Neurodata stored in the NWB format. MatNWB is used by both users and developers to create and interact with NWB and neuroscience tools to integrate with NWB. **[Source]**[59] **[Documentation]**[60]

NWBWidgets is an extensible library of widgets for visualizing NWB data in a Jupyter notebook (or lab). The widgets support navigation of the NWB file hierarchy and visualization of specific NWB data elements. **[Source]**[61]

NWB Schema defines the complete NWB data standard specification. The schema is a collection of YAML files in the NWB specification language describing all `neurodata_types` supported by NWB and their organization in an NWB file. **[Source]**[62] **[Documentation]**[63]

NWB Schema Language is a specialized variant of the HDMF schema language. The language includes minor modifications (e.g., use of the term `neurodata_type` instead of `data_type`) to make the language more intuitive for neuroscience users, but is otherwise identical to the HDMF schema language. Dedicated interfaces for creating and interacting with NWB schema are available in PyNWB. **[Documentation]**[64]

NWB Storage defines the mapping of NWB specification language primitives to HDF5 for storage of NWB files. **[Documentation]**[65]

Neurodata Extensions Catalog (NDX Catalog) is a community-led catalog of Neurodata Extensions (NDX) to the NWB data standard. All extensions mentioned in the text can be accessed directly via the catalog. **[Source]**[49] **[Online]**[66]

NWB Extensions Template (ndx-template) provides an easy-to-use template based on the Cookiecutter library for creating Neurodate Extensions (NDX) for the NWB data standard. **[Source]**[67]

NWB Staged Extensions is a repository for submitting new extensions to the NDX catalog **[Source]**[68]

DANDI

The DANDI archive was created by developing and integrating several opensource projects and BRAIN Initiative data standards (NWB, BIDS, NIDM). The Web browser application is built using the VueJS framework and the DANDI command line interface is built using Python and PyNWB. The initial backend of the archive was built on top of the Girder data management system, and is transitioning to a Django-based framework. The DANDI analysis hub is built using Jupyterhub deployed over a Kubernetes cluster. The different components of the archive are hosted on Amazon Web Services and the Heroku platform. The code repositories for the entire infrastructure are available on Github[40] under an Apache 2.0 license.

Acknowledgements: Research reported in this publication was supported by the following grants and institutions. OR, AT, RL, and KEB were supported by the National Institute Of Mental Health of the National Institutes of Health under Award Number R24MH116922 (PI: O. Ruebel) as well as through additional support by the Kavli Foundation (PI: KEB). BD and IS were supported by the BRAIN Initiative under award number U19-NS104590 to IS and by the Kavli Foundation. SG and BD were supported by the NIH BRAIN Initiative under award number 1R24MH117295-01A1. KS is supported by HHMI. Additional support for NWB was also provided by the Simons Foundation for the Global Brain grant 521921 to LF, and with additional funding from the Kavli Foundation and Simons Foundation for MatNWB to KS. KEB was additionally supported by the Weill Neurohub. CatalystNeuro is also supported by the Simons Foundation.

We thank the participants of the NWB hackathons and the broader NWB user and developer community for their feedback and contributions. We thank the current and former members of the NWB Executive Board: Kristofer Bouchard, Bing Brunton, Elizabeth Buffalo, Anne Churchland, Loren Frank, Satrajit Ghosh, Adam Kepecs, Lydia Ng, Huib Mansvelder, Ueli Rutishauser, Karel Svoboda, Christof Koch, Friedrich Sommer, Markus Meister, and Katrin Amunts.

Author Contributions:

Conceptualization: KEB, OR

Methodology: OR, AT, RL, SG, BKD

Software: RL, AT, OR, LN

Resource: KEB, OR, IS, LF, KS

Writing - Original Draft: KEB, BKD, SG, RL, OR, AT

Writing - Review & Editing: KEB, BKD, SG, RL, OR, AT, LF, IS, KS

Visualization: KEB, BKD, SG, RL, OR, AT

Supervision: KEB, OR

Project Administration: KEB, OR

Funding Acquisition: KEB, OR, SG, IS, LF, KS

Bibliography

1. Darwin, C. *The origin of species*. (PF Collier & son, 1909).
2. Kandel, E. R. *et al. Principles of Neural Science, Fifth Edition*. (McGraw Hill Professional, 2013).
3. Mallory, C. S. *et al.* Mouse entorhinal cortex encodes a diverse repertoire of self-motion signals. *Nat. Commun.* **12**, 671 (2021).
4. Kastner, D. B., Gillespie, A. K., Dayan, P. & Frank, L. M. Memory Alone Does Not Account for the Way Rats Learn a Simple Spatial Alternation Task. *J. Neurosci.* **40**, 7311–7317 (2020).
5. Bouchard, K. E., Mesgarani, N., Johnson, K. & Chang, E. F. Functional organization of human sensorimotor cortex for speech articulation. *Nature* **495**, 327–332 (2013).
6. Bezaire, M. J., Raikov, I., Burk, K., Vyas, D. & Soltesz, I. Interneuronal mechanisms of hippocampal theta oscillations in a full-scale model of the rodent CA1 circuit. *Elife* **5**, (2016).
7. Raikov, I. & Soltesz, I. Unpublished data.
8. Griffin, P. C. *et al.* Best practice data life cycle approaches for the life sciences. *F1000Res.* **6**, 1618 (2017).
9. Eaton, B. *et al.* NetCDF Climate and Forecast (CF) metadata conventions. (2003).
10. Hanisch, R. J. *et al.* Definition of the Flexible Image Transport System (FITS). *Astron. Astrophys. Suppl. Ser.* **376**, 359–380 (2001).
11. Brun, R. & Rademakers, F. ROOT—an object oriented data analysis framework. *Nucl. Instrum. Methods Phys. Res. A* **389**, 81–86 (1997).
12. Sejnowski, T. J., Churchland, P. S. & Movshon, J. A. Putting big data to good use in neuroscience. *Nat. Neurosci.* **17**, 1440–1441 (2014).
13. Bouchard, K. E. *et al.* High-Performance Computing in Neuroscience for Data-Driven Discovery, Integration, and Dissemination. *Neuron* **92**, 628–631 (2016).
14. Bouchard, K. E. *et al.* International Neuroscience Initiatives through the Lens of High-Performance Computing. *Computer* **51**, 50–59 (2018).
15. Teeters, J. L. *et al.* Neurodata Without Borders: Creating a Common Data Format for Neurophysiology. *Neuron* **88**, 629–634 (2015).
16. Martone, M. *et al.* NIX – Neuroscience information exchange format. *F1000Res.* **9**, (2020).
17. Tritt, A. J. *et al.* HDMF: Hierarchical Data Modeling Framework for Modern Science Data Standards. in *2019 IEEE International Conference on Big Data (Big Data)* 165–179 (2019).
18. Chandravadia, N. *et al.* A NWB-based dataset and processing pipeline of human single-neuron activity during a declarative memory task. *Sci Data* **7**, 78 (2020).

19. Ledochowitsch, P. *et al.* On the correspondence of electrical and optical physiology in in vivo population-scale two-photon calcium imaging. *Cold Spring Harbor Laboratory* 800102 (2019) doi:10.1101/800102.
20. Huang, L. *et al.* Relationship between simultaneously recorded spiking activity and fluorescence signal in GCaMP6 transgenic mice. *Cold Spring Harbor Laboratory* 788802 (2020) doi:10.1101/788802.
21. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016).
22. *MIES*. (Github).
23. Pologruto, T. A., Sabatini, B. L. & Svoboda, K. ScanImage: flexible software for operating laser scanning microscopes. *Biomed. Eng. Online* **2**, 13 (2003).
24. Siegle, J. H. *et al.* Open Ephys: an open-source, plugin-based platform for multichannel electrophysiology. *J. Neural Eng.* **14**, 045003 (2017).
25. Buccino, A. P. *et al.* SpikeInterface, a unified framework for spike sorting. *Cold Spring Harbor Laboratory* 796599 (2019) doi:10.1101/796599.
26. Allen Institute for Brain Science. IPFX. *Welcome to Intrinsic Physiology Feature Extractor (IPFX)* https://ipfx.readthedocs.io/.
27. Giovannucci, A. *et al.* CaImAn an open source tool for scalable calcium imaging data analysis. *Elife* **8**, (2019).
28. Pachitariu, M. *et al.* Suite2p: beyond 10,000 neurons with standard two-photon microscopy. *bioRxiv* 30 (2017) doi:10.1101/061507.
29. Ahanonu, B. *calciumImagingAnalysis (ciapkg): a software package for analyzing one- and two-photon calcium imaging datasets*. (2018). doi:10.5281/zenodo.2222295.
30. Dichter, B. K. *nwb-jupyter-widgets*. (Github).
31. Cantarelli, M. *et al. nwb-explorer*. (Github).
32. Magnotti, J. F., Wang, Z. & Beauchamp, M. S. RAVE: Comprehensive open-source software for reproducible analysis and visualization of intracranial EEG data. *Neuroimage* **223**, 117341 (2020).
33. Tauffer, L. & Dichter, B. *ecogVIS*. (Github).
34. Nasiotis, K. *et al.* Integrated open-source software for multiscale electrophysiology. *Sci Data* **6**, 231 (2019).
35. Garcia, S. *et al.* Neo: an object model for handling electrophysiology data in multiple formats. *Front. Neuroinform.* **8**, 10 (2014).
36. usc.edu, D. L. DABI. https://dabi.loni.usc.edu/home.

37. Gleeson, P. *et al.* Open Source Brain: A Collaborative Resource for Visualizing, Analyzing, Simulating, and Developing Standardized Models of Neurons and Circuits. *Neuron* **103**, 395–411.e5 (2019).
38. Dai, K. *et al.* The SONATA data format for efficient description of large-scale network models. *PLoS Comput. Biol.* **16**, e1007696 (2020).
39. Gorgolewski, K. J. *et al.* The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Sci Data* **3**, 160044 (2016).
40. *DANDI: Distributed Archives for Neurophysiology Data Integration*. (Github).
41. Bargmann, C. *et al.* BRAIN 2025: a scientific vision. *Brain Research through Advancing Innovative Neurotechnologies (BRAIN) Working Group Report to the Advisory Committee to the Director, NIH* (2014).
42. Holdgraf, C. *et al.* iEEG-BIDS, extending the Brain Imaging Data Structure specification to human intracranial electrophysiology. *Sci Data* **6**, 102 (2019).
43. *ndx-icephys-meta-record*. (Github).
44. Electrophysiology Stimulation Ontology Working Group. https://www.incf.org/sig/electrophysiology-stimulation-ontology-working-group.
45. Generator, M. Project information - NIH RePORTER - NIH research portfolio online reporting tools expenditures and results. https://projectreporter.nih.gov/project_info_description.cfm?aid=9795228.
46. Powell, K. The broken promise that undermines human genome research. *Nature* **590**, 198–201 (2021).
47. *Neurodata Without Borders*. (Github).
48. *Hierarchical Data Modeling Framework*. (Github).
49. *NWB Extension Catalog*. (Github).
50. *hdmf*. (Github).
51. The hierarchical data modeling framework — HDMF 2.3.0 documentation. https://hdmf.readthedocs.io.
52. HDMF. https://hdmf-dev.github.io/.
53. *hdmf-docutils*. (Github).
54. *hdmf-common-schema*. (Github).
55. Welcome to the HDMF-common format specification — HDMF-common specification v1.3.0 documentation. https://hdmf-common-schema.readthedocs.io.
56. *hdmf-schema-language*. (Github).
57. *pynwb*. (Github).

58. NWB for Python — PyNWB 1.4.0 documentation. https://pynwb.readthedocs.io.
59. *matnwb*. (Github).
60. matnwb. https://neurodatawithoutborders.github.io/matnwb/.
61. *nwb-jupyter-widgets*. (Github).
62. *nwb-schema*. (Github).
63. Welcome to the NWB format specification — NWB format specification v2.2.5 documentation. https://nwb-schema.readthedocs.io.
64. Welcome to the NWB specification language — NWB specification language v2.0.0-beta documentation. https://schema-language.readthedocs.io.
65. Welcome to NWB Storage — NWB Storage v1.0.0 documentation. https://nwb-storage.readthedocs.io.
66. NDX Catalog. https://nwb-extensions.github.io.
67. *ndx-template*. (Github).
68. *staged-extensions*. (Github).
69. Neurodata without borders – the kavli foundation. https://www.nwb.org/.
70. NWB Users Slack Workspace. https://join.slack.com/t/nwb-users/shared_invite/enQtNzMwOTcwNzQ2MDM5LWMyZDUwODJjYjM3MzMzYzZiNDk4ZTU3ZjQ3MmMxMmY5MDUyNzc0ZDI5ZjViYmJjYTQ5NjljOGFjZmMwOGIwZmQ.
71. NeurodataWithoutBorders (@NeurodataWB). *Twitter* https://twitter.com/NeurodataWB.
72. Neurodata Without Borders. https://www.youtube.com/channel/UCfD_mU-EFz135a9TpNFJP5A.
73. NWB Mailing List. https://mailchi.mp/fe2a9bc55a1a/nwb-signup.
74. Neurodata Without Borders: Neurophysiology (NWB:N). https://training.incf.org/collection/neurodata-without-borders-neurophysiology-nwbn.

Supplemental Material for
The Neurodata Without Borders ecosystem for neurophysiological data science

Supplemental Material 1. Intracellular Electrophysiology Example using NWB and DANDI

The following shows a simple example, demonstrating the use of NWB for storage of intracellular electrophysiology data. The file used in this example is from DANDISET 20 (available at https://dandiarchive.org/dandiset/000020) from the Allen Institute for Brain Science as part of the multimodal characterization of cell types in the mouse visual cortex (see https://portal.brain-map.org/explore/classes/multimodal-characterization/multimodal-characterization-mouse-visual-cortex).

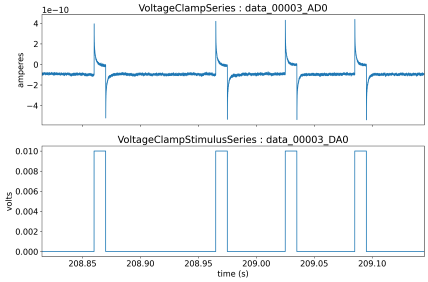
The following simple code example illustrates: 1) downloading of the file from DANDI, 2) reading the file with PyNWB, 3) visualization of the stimulus and response recording for a single sweep, and 4) visualization of the NWB file in NWB Widgets.

```
# import required libraries
import urllib
import os
import numpy as np
from matplotlib import pyplot as plt
import pynwb
from nwbwidgets import nwb2widget
from nwbwidgets.timeseries import show_indexed_timeseries_mpl

# Download the file from DANDI
url = 'https://girder.dandiarchive.org/api/v1/item/5f00fd6ba78c4b52bbb38e81/download'
filename = 'sub-1001658946_ses-1003020741_icephys.nwb'
res = urllib.request.urlretrieve(url, filename)

# Read the file with PyNWB
reader = pynwb.NWBHDF5IO(filename, mode='r', load_namespaces=True)
nwbfile = reader.read()

# Create a simple example visualization of the response and stimulus timeseries for a single sweep
# get the timeseries associated with a particular sweep number
sweep_number = 3
series = nwbfile.sweep_table.get_series(sweep_number)
# create a matplotlib figure for plotting
plt.rcParams['font.size'] = '16'
fig, (ax1, ax2) = plt.subplots(2, sharex=True, figsize=(12,8))
# plot the response and stimulus timeseries for the given sweep
show_indexed_timeseries_mpl(series[0], title=series[0].neurodata_type + " : " + series[0].name,
                            xlabel=None, ax=ax1)
show_indexed_timeseries_mpl(series[1], title=series[1].neurodata_type + " : " + series[1].name,  ax=ax2)
plt.show()
```



```
# Display the file with NWBWidgets
nwb2widget(nwbfile)
```

session_description:	PLACEHOLDER
identifier:	b215c33e0839253a6c9ee5d6d0a6975cb12437110acc9a7c26f59374de8ee055
session_start_time:	2020-01-27 21:53:18.318000+00:00
timestamps_reference_time:	2020-01-27 21:53:18.318000+00:00
session_id:	1003020741
institution:	Allen Institute for Brain Science
data_collection:	Free Memory: 14.6846 GB Please be patient while we export all existing acquired c…
source_script:	Release_2.0_20190602-977-g33b63e3b Date and time of last commit: 2020-05-29T…
source_script_file_name:	Vipr2-IRES2-Cre;Slc32a1-IRES2-FlpO;Ai65-508158.01.02.01

acquisition

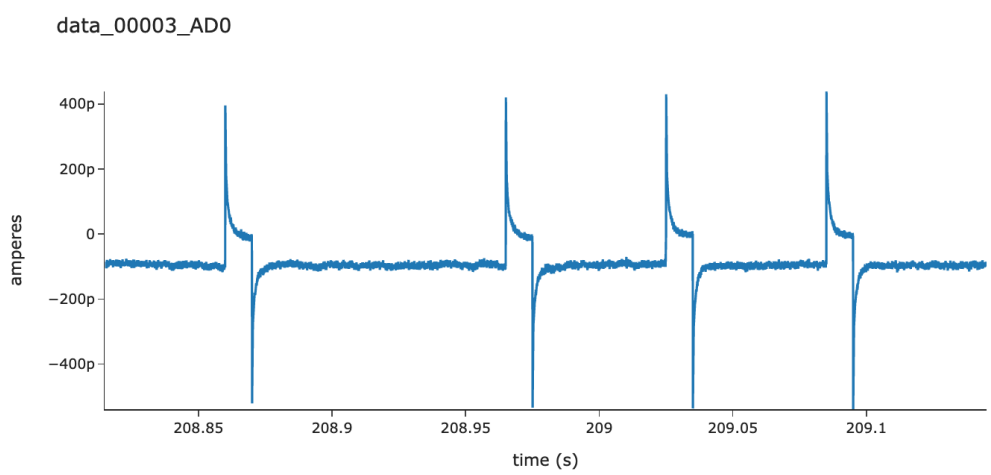
data_00000_AD0

data_00001_AD0

data_00002_AD0

data_00003_AD0

start (s) 208.82 duration (s): 0.3300



Supplementary Material 2. NWB Online and Social Media Resources

NWB online and social media resources provide additional resources for users and the broader community to engage with and learn about NWB. The nwb.org website[69] serves as the central entry-point for users to NWB and provides high-level information about NWB and links to all relevant online resources and tools discussed in the Methods. Additional online resources include **[Slack]**[70]**, [Twitter]**[71]**, [YouTube]**[72], and the **[NWB Mailing List]**[73].

Supplementary Material 3. NWB Training Resources

NWB provides a broad range of training resources for users and developers. Users who want to learn more about how to use NWB can view the NWB online video training course as part of the **[INCF Training Space]**[74]. Detailed code tutorials are further available as part of the **[PyNWB Documentation]**[58] and **[MatNWB Documentation]**[60].

For Neurodata Extensions (NDX), detailed documentation of versioning guidelines, sharing guidelines and strategies, and the proposal review process are available online as part of the NDX Catalog[66]. Step-by-step instructions for creating new NDX are provided as part of the **[NWB Extensions Template]**[67].

Additional resources for developers and data managers include the API documentation for **[HDMF]**[51]**, [PyNWB]**[58], and **[MatNWB]**[60] and documentation of the format schema as part of the **[NWB Schema]**[63], **[HDMF Common Schema]**[55]**, [NWB Storage]**[65]**,** and **[Specification Language]**[64].