Title

A modular architecture for organizing, processing and sharing neurophysiology data

Permalink
https://escholarship.org/uc/item/6rh1v3k2

Journal
Nature Methods, 20(3)

ISSN
1548-7091

Authors
Bonacchi, Niccolò
Chapuis, Gaelle A
Churchland, Anne K
et al.

Publication Date
2023-03-01

DOI
10.1038/s41592-022-01742-6

Peer reviewed

Published in final edited form as:
Nat Methods. 2023 March ; 20(3): 403–407. doi:10.1038/s41592-022-01742-6.

A modular architecture for organizing, processing and sharing neurophysiology data

The International Brain Laboratory,

Niccolò Bonacchi[1], **Gaelle Chapuis**[2,3], **Anne Churchland**[4], **Eric E. J. DeWitt**[1], **Mayo Faulkner**[2], **Kenneth D. Harris**[2], **Julia M. Huntenburg**[5], **Max Hunter**[2], **Inês Laranjeira**[1], **Cyrille Rossant**[2], **Maho Sasaki**[6], **Michael Schartner**[1], **Shan Shen**[6], **Nicholas A. Steinmetz**[7], **Edgar Y. Walker**[6], **Steven J. West**[8], **Olivier Winter**[1], **Miles Wells**[2]

[1]Champalimaud Center for the Unknown, Av. Brasília, 1400-038 Lisboa, Portugal

[2]Institute of Neurology, University College London, London WC1N 3BG, UK

Correspondence to: The International Brain Laboratory.

Correspondence: info+data@internationalbrainlab.org.

Author contributions

Niccolò Bonacchi [Conceptualization] (supporting). [Data curation] Data, metadata, and pipeline (equal). [Funding acquisition] (supporting). [Project administration] Meeting coordination (supporting), attendance (equal). [Resources] Computing and data storage (equal). [Software, Validation] Pipeline, core, quality control (equal), database and analysis libraries (supporting). [Visualization] Posters (lead), presentations (equal). [Writing – original draft] (equal). [Writing – review & editing] (equal).
Gaelle Chapuis [Data Curation] Helped with writing user guides detailing how to enter metadata in Alyx (supporting). [Software, Validation] Acted as naive user tester for ONE and DJ (supporting). [Project administration] Gathered and reported users' requirements (supporting).
Anne Churchland Contributed to project administration, funding acquisition, writing and revising.
Eric E. J. DeWitt [Investigation] (supporting) Contributed to analyses for example use of data [Writing] (supporting) Figures and draft text for example.
Mayo Faulkner. [Software] Contributed to implementation of backend data infrastructure and analysis libraries. [Data Curation] Contributed to curating datasets and assuring quality assurance.
Kenneth D. Harris. Contributed to design of overall data architecture, and to Alyx and ONE systems. Contributed to project administration, funding acquisition, writing and revising.
Julia Huntenburg. [Software] Contributed to implementation of backend data infrastructure, analysis libraries and continuous integration. [Data Curation] Contributed to dataset curation and quality assurance.
Michael Schartner. [Conceptualisation] Contributed to the development of dataset types related to video and their quality control metrics.
Max Hunter. Contributed to the design and development of Alyx.
Inês Laranjeira. [Investigation] Performed analyses for the example use case of the data architecture (lead). [Writing - original draft] (supporting)
Cyrille Rossant. Contributed to design and implementation of overall data architecture, and to Alyx and ONE systems.
Maho Sasaki Contributed to the design and implementation of the IBL Data Portal website.
Shan Shen Contributed to the design and implementation of the DataJoint pipeline and the IBL JupyterHub
Nicholas A. Steinmetz contributed to the design, testing, and development of Alyx, of dataset types, and of software tools for working with them.
Edgar Y. Walker Contributed to the design and implementation of the DataJoint pipeline and the IBL Data Portal.
Steven J. West contributed to the design of data structures for histological alignment.
Olivier Winter [Software] Implemented the backend data infrastructure. [Validation/Methodology] Designed full loop integration tests to allow maintenance of the codebase. [Data Curation] Fixed and updated erroneous datasets.
Miles Wells. [Software] Contributed to the design, testing and implementation of Alyx, its dataset types and of software tools that work with them. [Software, Validation] Contributed to the design and implementation of continuous integration and quality assurance systems. [Writing] Contributed to writing, reviewing and editing the text and appendices.

Competing Interests

The authors declare the following competing interests: E.Y.W. hold equity ownership in VathesLLC which provides development and consulting for the framework (DataJoint) described in this work. The remaining authors declare no competing interests

[3]Department of Basic Neuroscience, University of Geneva, Geneva, Switzerland

[4]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, USA

[5]Max Planck Institute for Biological Cybernetics, 72076 Tübingen, Germany

[6]DataJoint, Houston, TX 77027, USA

[7]Department of Biological Structure, University of Washington, Seattle, WA 98195, USA

[8]Sainsbury-Wellcome Centre, University College London, London WC1N 3BG, UK

Abstract

We describe an architecture for organizing, integrating, and sharing neurophysiology data in single labs or collaborations. It comprises a database linking data files to metadata and electronic lab notes; a module collecting data from multiple labs into one location; a protocol for searching and sharing data; and a module for automatic analyses which populates a website. These modules can be used together or individually, by single labs or worldwide collaborations.

Improving technology allows neurophysiologists to record ever larger datasets. The need for technologies to organize and share this data is growing as scientists begin to assemble into large, international teams. The International Brain Laboratory (IBL) is a collaboration studying the computations supporting decision-making[1]. We have developed modular data-management tools that enable individual labs and collaborations to:

- Manage experimental subject colonies and track subject- and experiment-level metadata
- Integrate data from multiple labs in a central store for sharing inside or outside the collaboration
- Access shared data through a simple programmatic interface
- Process incoming data through pipelines that automatically populate a website

Modern neurophysiological datasets comprise multiple recordings from multiple subjects, recorded using diverse devices. These data must be preprocessed, time-aligned, and integrated with data such as locations of recording electrodes before they can be used to draw scientific conclusions[2–8]. Distributed collaborations pose distinct challenges: while public data release must wait for careful quality control, scientists within the collaboration require immediate access to specific data. This store must be searchable and allow downloading and also revision of individual items, because preprocessing and quality control methods are still evolving[9–11].

We addressed these problems with an architecture consisting of four modules (Figure 1). The first module is a Web interface for colony management and electronic lab notebook, that links files arising from each experiment to relevant metadata. The second module integrates data from multiple labs into a central database and bulk data store, providing immediate access while allowing updates of individual items. The third automatically runs analyses on newly-arrived data, providing results via a Web interface.

The fourth allows standardization, access and sharing of the data. Full documentation can be found at https://docs.internationalbrainlab.org/ and through the links at https://www.internationalbrainlab.com/tools.

To manage data within each lab, we developed “Alyx”: a relational database that links colony management, metadata, and lab notes to experimental data files. A web GUI allows users to enter metadata as it arrives (such as birth, weaning, genotyping, surgeries or experiments), and a REST API allows experiment control software to automatically enter metadata with a one-line command. Bulk data files are stored on a lab server, and linked to experiment and subject metadata in the database. This tool can be used by single labs as well as collaborations: it was developed in one member lab prior to IBL’s founding, and is now used by several labs worldwide for non-IBL work. An Alyx user guide can be found here, or linked via our main documentation page.

Integrating data between labs raises challenges of size and complexity. Large-scale electrophysiology produces hundreds of gigabytes per experiment, for which we have designed a novel 3-fold lossless compression algorithm (Appendix 1). A single IBL experiment generates over 150 raw and processed data files. We have devised conventions for organizing and naming these files, termed the “Open Neurophysiology Environment” (ONE; Appendix 2; https://int-brain-lab.github.io/ONE/), which formalizes how to encode cross-references between files, time synchronization, and versioning, and allows local and remote access via an API. ONE provides a simple way to standardize and share data from individual labs, by specifying standard filenames for common data types (Appendix 3) and defining conventions for naming lab-specific data files. Files from multiple labs are integrated by uploading nightly from lab servers to a central server using Globus Online[15], coordinated by a central Alyx database which also stores metadata from all labs.

Neurophysiology data requires preprocessing, such as spike sorting and video analysis. We developed a task management system that uses computers in member labs as a processing pool. Computers query the Alyx database for a list of outstanding preprocessing tasks, determined by a dependency graph. Because Alyxis accessed through http, this works despite different universities’ diverse firewall policies, and allows monitoring, logging, and restarting all preprocessing tasks. Higher-level analyses are automatically run on newly preprocessed data using DataJoint[14], which runs automated analyses and places the results on a website, including summaries of behavioral performance allowing scientists to monitor training progress, and basic analyses of spike trains. While manual curation of the full dataset will be required before public release, an illustrative curated subset of these data are available on a public website (https://data.internationalbrainlab.org).

To access data, an API allows users to search experiments and load data from the ONE files directly into Python (Appendix 3). This API allows both collaborations and individual labs to share data using the same standard. A large collaboration such as IBL can host files on a server such as AWS, and run an Alyxserver which allows users to rapidly search and selectively download the data. Individual labs can release data compatible with the same API by “uploading and forgetting” a zip of ONE files for users to download in toto (instructions here). Users can also access data via Neurodata Without Borders (NWB)[12,13] using software that translates from the ONE standard (https://github.com/catalystneuro/IBL-to-nwb; Supplementary Table 1), or through DataJoint[14]. A comparison of these and other sharing systems is in Appendix 4. The analyses in a recently-published paper[1] were made using this system, and an additional example is provided in Appendix 5.

The IBL architecture was designed for our large-scale collaboration, but its modular design allows components to be used by individual labs and smaller-scale collaborations. The Alyx system provides easy-to-use colony management and electronic lab notebook features for labs or collaborations, linking experimental files to this metadata. The ONE conventions allow data to be organized within a lab and shared externally, using standards that scale to large collaborations. Larger collaborations can also benefit from other features such as the DataJoint architecture to perform automated analyses for web display. We hope that these tools, and additional software we have provided (Appendix 6), will help pave the way forward to an era in which data from neurophysiology labs is integrated and shared on a routine basis.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the Wellcome Trust (209558 to IBL, 216324 to IBL) and Simons Foundation (to IBL).

Data availability

Data for Figures 2 and 3 is available at https://data.internationalbrainlab.org/.

Code availability

All code described in this manuscript is freely available and is listed in Supplementary Table 1 along with links to their respective repositories. The behavior data were collected using Bonsai and pyBpod, available at https://github.com/int-brain-lab/iblrig. Meta data were stored in a custom database available at https://github.com/cortex-lab/alyx. The data were processed using the custom data pipelines ibllib (https://github.com/int-brain-lab/iblrig) and DataJoint (https://datajoint.io/). The data were accessed using ONE (https://github.com/int-brain-lab/ONE) and DataJoint (https://github.com/int-brain-lab/IBL-pipeline).

References

1. The International Brain Laboratory. et al. Standardized and reproducible measurement of decision-making in mice. eLife. 2021; 10 e63711 [PubMed: 34011433]

2. Pachitariu M, et al. Suite2p: beyond 10,000 neurons with standard two-photon microscopy. bioRxiv. 2017; 061507 doi: 10.1101/061507

3. Mathis A, et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. Nat Neurosci. 2018; 21: 1281–1289. [PubMed: 30127430]

4. Giovannucci A, et al. CaImAn an open source tool for scalable calcium imaging data analysis. eLife. 2019; 8 e38173 [PubMed: 30652683]

5. Vogelstein JT, et al. Fast nonnegative deconvolution for spike train inference from population calcium imaging. J Neurophysiol. 2010; 104: 3691–704. [PubMed: 20554834]

6. Pachitariu, M, Steinmetz, NA, Kadir, SN, Carandini, M, Harris, KD. Advances in Neural Information Processing Systems. Lee, DD, Sugiyama, M, Luxburg, UV, Guyon, I, Garnett, R, editors. Vol. 29. Curran Associates, Inc; 2016. 4448–4456.

7. Wiltschko AB, et al. Revealing the structure of pharmacobehavioral space through motion sequencing. Nat Neurosci. 2020; 23: 1433–1443. [PubMed: 32958923]

8. Vogelstein JT, et al. Discovery of Brainwide Neural-Behavioral Maps via Multiscale Unsupervised Structure Learning. Science. 2014; 344: 386–392. [PubMed: 24674869]

9. Siegle JH, et al. Survey of spiking in the mouse visual system reveals functional hierarchy. Nature. 2021; 592: 86–92. [PubMed: 33473216]

10. Hill DN, Mehta SB, Kleinfeld D. Quality metrics to accompany spike sorting of extracellular signals. J Neurosci. 2011; 31: 8699–705. [PubMed: 21677152]

11. Harris KD, Quiroga RQ, Freeman J, Smith SL. Improving data quality in neuronal population recordings. Nat Neurosci. 2016; 19: 1165–1174. [PubMed: 27571195]

12. Teeters JL, et al. Neurodata Without Borders: Creating a Common Data Format for Neurophysiology. Neuron. 2015; 88: 629–634. [PubMed: 26590340]

13. Rübel O, et al. The Neurodata Without Borders ecosystem for neurophysiological data science. bioRxiv. 2021; 2021.03.13.435173 doi: 10.1101/2021.03.13.435173

14. Yatsenko D, et al. DataJoint: managing big scientific data using MATLAB or Python. bioRxiv. 2015; 031658 doi: 10.1101/031658

15. Foster I. Globus Online: Accelerating and Democratizing Science through Cloud-Based Services. IEEE Internet Comput. 2011; 15: 70–73.

16. Wang Q, et al. The Allen Mouse Brain Common Coordinate Framework: A 3D Reference Atlas. Cell. 2020; 181: 936–953. e20 [PubMed: 32386544]

17. Allen Institute for Brain Science. Allen Mouse Brain Atlas (2015) with region annotations. 2017.

18. Urai AE, et al. Citric Acid Water as an Alternative to Water Restriction for High-Yield Mouse Behavior. eNeuro. 2021; 11 (1) 8.

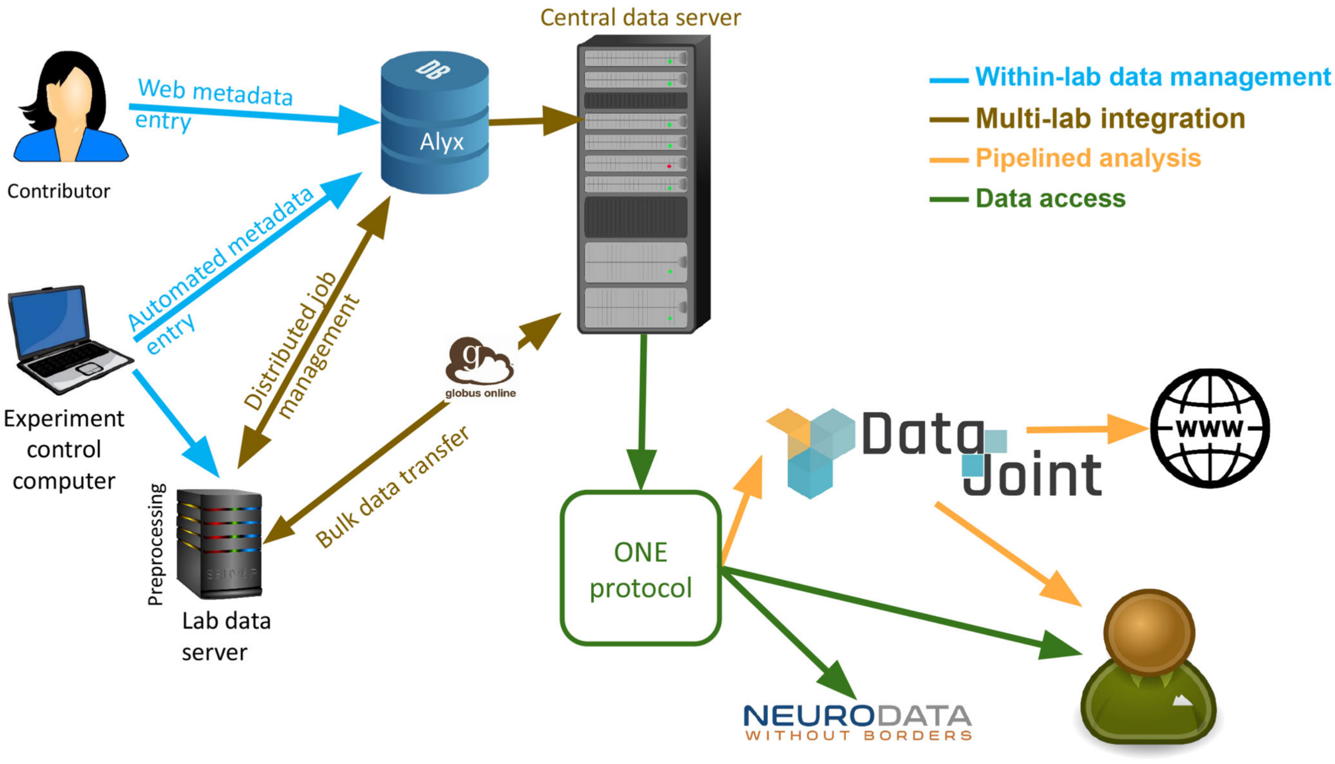

Figure 1.

IBL data architecture. The “Alyx” database links colony management and electronic lab notebook metadata to experimental data files on a lab data server. Data from multiple labs are integrated on a central server, and distributed job management coordinates pre-processing on lab servers. Data are accessed via the Open Neurophysiology Environment (ONE) protocol, with adaptors for Neurodata Without Borders (NWB)[12,13] and DataJoint[14], which also performs pipelined analyses for automatic display on a website.