# Automating Data Integration and Publishing for Neuroimaging via LSLAutoBIDS

Manpa Barman[1], Jan Range[2,3], and Benedikt Ehinger[1,3]

[1]University of Stuttgart, Institute for Visualization and Interactive Systems
[2]University of Stuttgart, Institute of Biochemistry
[3]University of Stuttgart, Stuttgart Center for Simulation Science

February 26, 2026

**Abstract**

Cognitive neuroscience routinely collect large datasets, yet data integration, curation, version control, and publishing is rarely automated. Here, we present such a workflow, offering open science by design, with an implementation in the python based `LSLAutoBIDS` open-source package. We first describe our exemplary workflow based on LabStreamingLayer (to integrate), BIDS (to transform), DataLad (to version), and Dataverse (to publish), before discussing the place such tools can have in future data collection efforts.

## Keywords:

BIDS, EEG, Lab Streaming Layer, Dataverse, Datalad, data collection, datasets, open science, data standards

# 1 Introduction

Modern neuroscience research often relies on collecting complex, multimodal datasets, for example, combining behavioural, eye-tracking, and neuroimaging (e.g., EEG) data. Such datasets are valuable for understanding cognitive processes, but are often challenging to collect due to the resource-intensive nature of experiments and the logistical demands. They are often collected by trainees or research assistants with limited prior experience or training. And even to more experienced data collectors, the following examples of what can go wrong will sound very familiar: 1) the same participant ID was used twice in a row, and data was irreversibly overwritten. 2) While running a study, the experiment was modified to fix a bug, but it was not documented after which subject it was fixed - later it remained unclear whether early subjects data were retrogradely fixed or not. 3) A masters student left the lab without documenting their last dataset, leaving a mess of various files on various computers. 4) To move files from the various recording computers to their analysis system, every lab user seems to have their own idiosyncratic workflow, from USB sticks to WiFi hotspots. In this paper, we show that these issues can be prevented while at the same time saving time and training effort.

To increase the robustness of data recording pipelines, as well as to encourage data sharing, we present an approach to automate parts of this workflow. The automation we propose can be understood under the umbrella of open science by design[15]. The central idea is to change infrastructure and workflows in a way that they are open from the point of creation, making open science a by-product and not an afterthought[15]. In our case, the new data is curated and inserted into an existing dataset, versioned, and (privately) archived in a remote repository immediately after the recording. Additionally, many open data check boxes are also immediately marked without additional effort by the experimenter.

Our proposed workflow automates four main aspects of such data recording pipelines: data integration, data curation, data versioning, and data publishing.

**Data Integration:** Collected data often comes from many sources. LabStreamingLayer (LSL) [13] has become a standard way to record multiple data streams and synchronize their time series. In addition, other data sources need to be collected from devices that cannot send them via LSL (e.g., experimental logging files, proprietary eye-tracking data).

**Data Curation:** To allow machines and humans to navigate such a diverse data collection, we want to structure them. The Brain Imaging Data Structure (BIDS) [7] standard provides a well-described data curation environment, offering consistent directory structure, file naming conventions, and metadata descriptors for neuroimaging data. Once data is structured in BIDS format, the dataset can be seamlessly integrated into BIDS-compliant analysis pipelines.

**Data Versioning:** The next aspect is to include datasets into version control. While much progress has been made in version control for analysis code, datasets are only rarely versioned by default. DataLad [9] combines git and git-annex, allowing for efficient versioning of not only code, but also binary files. Crucially, adding new subjects, annotations, or corrections will leave a version trace, making the data collection robust to accidentally deleting or overwriting data and improving transparency of data provenance [10].

**Data Publishing:** Our last step, which traditionally is often only taken after the data has been fully analyzed, is the long-term archiving and (access-restricted) publishing. Immediately publishing the raw data has the benefit of frontloading (and externalizing) the additional effort required to publish the data. Automating this is made possible by linking the versioned DataLad [9] datasets directly with FAIR-enabling [20] data repositories such as Dataverse [5], which will provide Digital Object Identifiers (DOI), customizable access controls, archival service, and further automatic versioning.

In our work, we addressed all four aspects in a single automated workflow. Our contributions are: (1) we discuss workflow to integrate community standards for data curation, automated publishing, and version control, with the potential to generalize across modalities and experimental paradigms, (2) we introduce `LSLAutoBIDS`, a Python package as an example implementation for such a workflow, using Lab Streaming Layer [13], converting to BIDS, versioning via Datalad [9], and publishing via Dataverse.

# 2 Related Works

The pioneering work of Dobson et al. [6] should be highlighted: They present ReproIn, a software package quite similar to ours, which converts DICOM data from an MR scanner to BIDS, and finally saves it in a DataLad repository. Compared to our implementation, ReproIn is optimized to work with different MR scanners, whereas we use LSL to integrate time-series data more directly. Instead of `heudiconv`, we use `mnelab`, `pylsl`, and `mne-bids` for the data curation stage. While ReproIn does not automatically link the DataLad repository to a Dataverse, this additional step could be readily implemented for ReproIn as well.

Another pioneering approach is taken at the Donders Institute for Brain and Cognition [1]. There, all fMRI and MEG data are automatically archived in a data acquisition collection (DAC). Other data needs to be added manually to the DAC. While the DAC is centralized, it is typically only used internally, and later, a data sharing

collection (DSC) is created, which can be publicly shared.

A different example for a data curation workflow, incorporating standardization and publishing, i.e, from raw data acquisition over standardization to analysis, is presented in Stawiski et al., [19]. In this project, they collected heterogeneous raw data of more than 100 participants in two clinics, transformed them to BIDS, and used SQLite to manage the resulting dataset. Even though they do not discuss versioning and publishing, it is easily conceivable how these steps could be added to their workflow.

The previously introduced **Brain Imaging Data Structure** [7], originally proposed for magnetic resonance imaging (MRI), is a community standard for data curation and sharing of brain data within research communities. This standard is designed using the FAIR (findability, accessibility, interoperability, and reusability) principles [20], contributing to efficient scientific data curation and stewardship. Following the release of the BIDS standard, numerous extensions were developed, concerning integration of BIDS with different neuroimaging techniques like magnetoencephalography (MEG) [16], intracranial electroencephalography (iEEG) [11], electroencephalography (EEG) [18], Positron Emission Tomography (PET) [17], and others. Data sharing and reuse of these BIDS-complaint datasets is even further simplified using open-source sharing platforms, for instance, OpenNeuro [14], which currently hosts more than 600 datasets in compliance with BIDS.

To **convert datasets** to be BIDS compliant, packages like `mne-bids` [3], [21] or `BIDSCoin` [22] (using the sova2coin plugin) are essential to move and convert source data to BIDS-compliant datasets, and subsequently, the BIDS validator [7] can be used to validate the data collection. To make the data conversion even more user-friendly for researchers, BIDSCoin [22] offers a graphical user interface, making it accessible even to those without programming experience. As far as we know, there is no dedicated LSL-to-BIDS converter available.

Unlike software and code versioning, **data versioning** is more rarely practiced, given issues in practicality to version large binary files. Nevertheless, version control is an important aspect in data collection to ensure data traceability and an error-free workflow. DataLad [9], based on git-annex, nicely extends the capabilities of classical git versioning tools to large, binary datasets. DataLad is actively used in hundreds of studies and underlies all datasets on OpenNeuro.

# 3 LSLAutoBIDS

Figure 1: Flowchart of LSLAutoBIDS. We start with different data streams recorded using LabStreamingLayer and as other proprietary file formats. (1) We use LSLAutoBIDS to integrate data that is collected across different computers and synchronise the data streams temporally. (2) Next, we organize these data into the BIDS structure. (3) Using DataLad, we version them, and (4) finally upload them to an open repository like DataVerse or potentially OpenNeuro.

`LSLAutoBIDS` is an open-source Python package developed and actively used by the Computational Cognitive Science Lab at the University of Stuttgart. It offers a modular and reproducible workflow tailored for studies using LSL based data acquisition, specifically targeting the integration of EEG and eye-tracking modalities.

In this setup, participant-level EEG data streams, but also other LSL streams (e.g., motion tracking, mobile eye-tracking), are recorded using the LSL protocol [12], which allows for sub-millisecond time-synchronization of heterogeneously sampled streams and results in an XDF container file, most commonly via the `LSLRecorder` [2] package. Project metadata like authors, license, experimental description, but also dataverse details are specified in one central configuration-toml file, which is then used to retrieve these project-specific metadata during the conversion process. The recorded raw data streams are then converted into the Brain Imaging Data Structure (BIDS) [7] standard using the `mnelab` [4], `pylsl`[12], `mne`[8], and `mne_bids` [3] packages. The converted EEG data are then validated with the `BIDSValidator`, while non-BIDS files like project metadata, configuration files, and additional modalities such as eye-tracking data are listed in a `.bidsignore` file to exclude them from validation. Thus, we are compliant with the `BIDSValidator` for our output. Once validated, the BIDS-compliant dataset is automatically deposited into a Dataverse repository, along with the experiment stimulus files and the raw source data.

In addition to LSL-based EEG data acquisition, `LSLAutoBIDS` supports the incorporation of non-LSL data. For example, in our use case, eye-tracking data are collected using the EyeLink 1000 Plus eye tracker simultaneously with EEG data collection, which produces proprietary Eyelink data format (EDF) files that cannot easily be streamed via LSL. Other examples are an electronic lab notebook record, log files of the experiment, and a compressed archive of the experimental code used in that individual session. `LSLAutoBIDS` accommodates these files as a secondary data modality currently implemented via user-defined folders, filename specifications, and regular expression matching, and is able to organize them within the appropriate BIDS subdirectories to publish them alongside EEG data.

Version control is integrated into the pipeline using DataLad [9], enabling precise tracking of all data and metadata changes across the research lifecycle, including modification or re-transformation of the individual

files. Datalad is very powerful, but we use only the basic functionalities to simplify adoption for new users. We are not using the `datalad run` or `datalad rerun` capabilities, but are only concerned with the versioning of large files (`datalad save` and `datalad push`). The two minimal commands a user needs to know are `datalad clone` to checkout a repository and `datalad get` to actually download large files. This split between providing only symbolic links to all binary files first and only then explicitly downloading, e.g., a subset of the dataset, allows for easy exploration of terabyte-sized datasets.

After each recording session, the dataset is then uploaded to an initially private Dataverse repository, where it is persistently archived with the specified metadata under a DOI. After data collection is finished, the dataset can be publicly released and shared, potentially under a data sharing agreement.

The architecture of `LSLAutoBIDS` is deliberately designed to be extensible. While we developed this tool with our setup in mind (Eyelink1000, ANT Neuro EEG, VR-setups), we will continuously add new devices and allow for further customizability. Already now, additional files, such as behavioral data, audio recordings, or physiological signals, can be integrated by extending configuration templates and adding corresponding processing steps. This generalizability makes `LSLAutoBIDS` not just a tool for EEG and eye-tracking studies, but a proof of concept for broader multimodal data workflows in cognitive neuroscience.

How do workflows like LSLAutoBIDS address the examples raised in the introduction? 1) If data is immediately versioned and archived, then overwritten data can always be recovered. 2) As experimental files are included in the archive for each subject, it is easy to find any changes in the experimental code during data recording. 3) Even if a student leaves without further documenting the dataset, it is already in a cleaned up state, including metadata, greatly simplifying further use of the data. 4) Training of new users is greatly simplified, manual copying data via e.g. USB sticks and other idiosyncratic workflows are no longer necessary.

# 4 Conclusion

Practicing open science by design and frontloading the conversion to a citable data publication efficiently addresses many hurdles researchers face in data collection and publishing. Not only is any recorded dataset immediately converted to BIDS, it also is archived, findable, shareable, backed up, and versioned.

Whether the final dataset is publicly available or only used internally (e.g., due to privacy issues in non-defaced MRIs), we think such a workflow will be helpful to many laboratories.

# 5 Code and Data Availability

The LSLAutoBIDS package is continuously developed and freely available: `https://github.com/s-ccs/LSLAutoBIDS` or via Zenodo `https://zenodo.org/records/15525822`.

# 6 Conflict of Interest

The authors declare no conflicts of interest that may bias or could be perceived to bias this work.

# 7 Funding

Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundations) in the Emmy Noether Programme - Project-ID 538578433 - and Germanys Excellence Strategy EXC 2075 - 390740016.

# References

[1] The RDM handbook — DCCN RDM handbook documentation.

[2] labstreaminglayer/app-LabRecorder, 2025-11-20. original-date: 2018-02-28T12:17:09Z.

[3] Stefan Appelhoff, Matthew Sanderson, Teon L. Brooks, Marijn van Vliet, Romain Quentin, Chris Holdgraf, Maximilien Chaumon, Ezequiel Mikulan, Kambiz Tavabi, Richard Höchenberger, Dominik Welke, Clemens Brunner, Alexander P. Rockhill, Eric Larson, Alexandre Gramfort, and Mainak Jas. MNE-BIDS: Organizing electrophysiological data into the BIDS format and facilitating their analysis. *Journal of Open Source Software*, 4(44):1896, 2019-12-18.

[4] Clemens Brunner. MNELAB: a graphical user interface for MNE-python. *Journal of Open Source Software*, 7(78):4650, 2022-10-10.

[5] Merce Crosas. The dataverse network: An open-source application for sharing, discovering and preserving data. *D-Lib Magazine*, Volume 17, 2011.

[6] J. Dobson, T. Sackett, Chandana Kodiweera, J. Haxby, M. Goncalves, Satrajit S. Ghosh, and Y. Halchenko. Presentations 2000 : ReproIn : automatic generation of shareable , version-controlled BIDS datasets from MR scanners. 2018.

[7] Krzysztof J. Gorgolewski, Tibor Auer, Vince D. Calhoun, R. Cameron Craddock, Samir Das, Eugene P. Duff, Guillaume Flandin, Satrajit S. Ghosh, Tristan Glatard, Yaroslav O. Halchenko, Daniel A. Handwerker, Michael Hanke, David Keator, Xiangrui Li, Zachary Michael, Camille Maumet, B. Nolan Nichols, Thomas E. Nichols, John Pellman, Jean-Baptiste Poline, Ariel Rokem, Gunnar Schaefer, Vanessa Sochat, William Triplett, Jessica A. Turner, Gaël Varoquaux, and Russell A. Poldrack. The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, 3(1):160044, 2016-06-21. Publisher: Nature Publishing Group.

[8] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti S. Hämäläinen. MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience*, 7(267):1–13, 2013.

[9] Yaroslav O. Halchenko, Kyle Meyer, Benjamin Poldrack, Debanjum Singh Solanky, Adina S. Wagner, Jason Gors, Dave MacFarlane, Dorian Pustina, Vanessa Sochat, Satrajit S. Ghosh, Christian Mönch, Christopher J. Markiewicz, Laura Waite, Ilya Shlyakhter, Alejandro de la Vega, Soichi Hayashi, Christian Olaf Häusler, Jean-Baptiste Poline, Tobias Kadelka, Kusti Skytén, Dorota Jarecka, David Kennedy, Ted Strauss, Matt Cieslak, Peter Vavra, Horea-Ioan Ioanas, Robin Schneider, Mika Pflüger, James V. Haxby, Simon B. Eickhoff, and Michael Hanke. DataLad: distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software*, 6(63):3262, 2021-07-01.

[10] Michael Hanke, Franco Pestilli, Adina S. Wagner, Christopher J. Markiewicz, Jean-Baptiste Poline, and Yaroslav O. Halchenko. In defense of decentralized research data management. *Neuroforum*, 27(1):17–25, 2021-02-01. Publisher: De Gruyter.

[11] Christopher Holdgraf, Stefan Appelhoff, Stephan Bickel, Kristofer Bouchard, Sasha D'Ambrosio, Olivier David, Orrin Devinsky, Benjamin Dichter, Adeen Flinker, Brett L. Foster, Krzysztof J. Gorgolewski, Iris Groen, David Groppe, Aysegul Gunduz, Liberty Hamilton, Christopher J. Honey, Mainak Jas, Robert Knight, Jean-Philippe Lachaux, Jonathan C. Lau, Christopher Lee-Messer, Brian N. Lundstrom, Kai J. Miller, Jeffrey G. Ojemann, Robert Oostenveld, Natalia Petridou, Gio Piantoni, Andrea Pigorini, Nader Pouratian, Nick F. Ramsey, Arjen Stolk, Nicole C. Swann, François Tadel, Bradley Voytek, Brian A. Wandell, Jonathan Winawer, Kirstie Whitaker, Lyuba Zehl, and Dora Hermes. iEEG-BIDS, extending the brain imaging data structure specification to human intracranial electrophysiology. *Scientific Data*, 6(1):102, 2019-06-25. Publisher: Nature Publishing Group.

[12] Christian Kothe. chkothe/pylsl, 2025-05-18. original-date: 2015-04-25T04:00:38Z.

[13] Christian Kothe, Seyed Yahya Shirazi, Tristan Stenner, David Medine, Chadwick Boulay, Matthew I. Grivich, Tim Mullen, Arnaud Delorme, and Scott Makeig. The lab streaming layer for synchronized multimodal recording, 2024-02-14. Pages: 2024.02.13.580071 Section: New Results.

[14] Christopher J. Markiewicz, Krzysztof J. Gorgolewski, Franklin Feingold, Ross Blair, Yaroslav O. Halchenko, Eric Miller, Nell Hardcastle, Joe Wexler, Oscar Esteban, Mathias Goncalves, Anita Jwa, and Russell A. Poldrack. OpenNeuro: An open resource for sharing of neuroimaging data. 2021-06-29. Repository: Neuroscience.

[15] National Academies of Sciences, Engineering, and Medicine, Policy and Global Affairs, Board on Research Data and Information, and Committee on Toward an Open Science Enterprise. *Open science by design.* National Academies Press, 2018-08.

[16] Guiomar Niso, Krzysztof J. Gorgolewski, Elizabeth Bock, Teon L. Brooks, Guillaume Flandin, Alexandre Gramfort, Richard N. Henson, Mainak Jas, Vladimir Litvak, Jeremy T. Moreau, Robert Oostenveld, Jan-Mathijs Schoffelen, Francois Tadel, Joseph Wexler, and Sylvain Baillet. MEG-BIDS, the brain imaging data structure extended to magnetoencephalography. *Scientific Data*, 5(1):180110, 2018-06-19. Publisher: Nature Publishing Group.

[17] Martin Norgaard, Granville J. Matheson, Hanne D. Hansen, Adam Thomas, Graham Searle, Gaia Rizzo, Mattia Veronese, Alessio Giacomel, Maqsood Yaqub, Matteo Tonietto, Thomas Funck, Ashley Gillman, Hugo Boniface, Alexandre Routier, Jelle R. Dalenberg, Tobey Betthauser, Franklin Feingold, Christopher J.

Markiewicz, Krzysztof J. Gorgolewski, Ross W. Blair, Stefan Appelhoff, Remi Gau, Taylor Salo, Guiomar Niso, Cyril Pernet, Christophe Phillips, Robert Oostenveld, Jean-Dominique Gallezot, Richard E. Carson, Gitte M. Knudsen, Robert B. Innis, and Melanie Ganz. PET-BIDS, an extension to the brain imaging data structure for positron emission tomography. *Scientific Data*, 9(1):65, 2022-03-02. Publisher: Nature Publishing Group.

[18] Cyril R. Pernet, Stefan Appelhoff, Krzysztof J. Gorgolewski, Guillaume Flandin, Christophe Phillips, Arnaud Delorme, and Robert Oostenveld. EEG-BIDS, an extension to the brain imaging data structure for electroencephalography. *Scientific Data*, 6(1):103, 2019-06-25. Publisher: Nature Publishing Group.

[19] Marc Stawiski, Vittoria Bucciarelli, Dorian Vogel, and Simone Hemm. Optimizing neuroscience data management by combining REDCap, BIDS and SQLite: a case study in deep brain stimulation. *Frontiers in Neuroinformatics*, 18:1435971, 2024-09-05.

[20] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino Da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 'T Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene Van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan Van Der Lei, Erik Van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1):160018, 2016-03-15.

[21] Mantilla-Ramos Yorguin-Jose, Hoyos-Madera Brayan-Andres, Bollmann Steffen, Narayanan Aswin, White David, Civier Oren, and Johnstone Tom. SOVABIDS: EEG-to-BIDS conversion software focused on automation, reproducibility and interoperability, 2025-03-20.

[22] Marcel Peter Zwiers, Stefano Moia, and Robert Oostenveld. BIDScoin: A user-friendly application to convert source data to brain imaging data structure. *Frontiers in Neuroinformatics*, 15, 2022-01-13. Publisher: Frontiers.