The Journal of Open Source Software

HeuDiConv — flexible DICOM conversion into structured directory layouts

Yaroslav O. Halchenko [1], **Mathias Goncalves** [2], **Satrajit Ghosh** [3], **Pablo Velasco** [4], **Matteo Visconti di Oleggio Castello** [5], **Taylor Salo** [6], **John T. Wodder II** [1], **Michael Hanke** [7,8], **Patrick Sadil** [22], **Krzysztof Jacek Gorgolewski** [24], **Horea-Ioan Ioanas** [1], **Chris Rorden** [9], **Timothy J. Hendrickson** [10,11], **Michael Dayan** [12], **Sean Dae Houlihan** [1,13], **James Kent** [14], **Ted Strauss** [15], **John Lee** [16], **Isaac To** [1], **Christopher J. Markiewicz** [2], **Darren Lukas** [17], **Ellyn Butler** [23], **Todd Thompson** [13], **Maite Termenon** [18,19], **David V. Smith** [20], **Austin Macdonald** [1], and **David N. Kennedy** [21]

1 Center for Open Neuroscience, Department of Psychological and Brain Sciences, Dartmouth College, Hanover, NH, USA **2** Department of Psychology, Stanford University, CA, USA **3** McGovern Institute, Massachusetts Institute of Technology, Cambridge, MA, USA **4** Flywheel Exchange LLC, Minneapolis, MN, USA **5** University of California, Berkeley, Berkeley, CA, USA **6** Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA **7** Institute of Neuroscience and Medicine, Brain & Behaviour (INM-7), Research Center Jülich, Jülich, Germany **8** Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany **9** Department of Psychology, University of South Carolina, Columbia, SC, USA **10** Masonic Institute for the Developing Brain, University of Minnesota, Minneapolis, MN, USA **11** Minnesota Supercomputing Institute, University of Minnesota, Minneapolis, MN, USA **12** Human Neuroscience Platform, Fondation Campus Biotech Geneva, Geneva, Switzerland **13** Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA **14** Department of Psychology, University of Texas at Austin, Austin, TX, USA **15** McConnell Brain Imaging Centre, McGill University, Montreal, QC, Canada **16** Data Science and Sharing Team, National Institute of Mental Health, Bethesda, MD, USA **17** Institute for Glycomics, Griffith University, QLD, Australia **18** Biomedical Engineering Department, Faculty of Engineering, Mondragon University, Mondragon, Spain **19** BCBL, Basque center on Cognition, Brain and Language, San Sebastian, Spain **20** Department of Psychology and Neuroscience, Temple University, Philadelphia, PA, USA **21** Departments of Psychiatry and Radiology, University of Massachusetts Chan Medical School, Worcester, MA, USA **22** Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA **23** Department of Psychology, Northwestern University, Evanston, IL, USA **24** Emeritus of Department of Psychology, Stanford University, CA, USA

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Britta Westner

Reviewers:

- @marcelzwiers
- @JLefortBesnard

Submitted: 06 July 2023
Published: unpublished

Summary

In order to support efficient processing, data must be formatted according to standards prevalent in the field, and widely supported among actively developed analysis tools. The Brain Imaging Data Structure (BIDS) (K. J. Gorgolewski et al., 2016) is an open standard designed for computational accessibility, operator legibility, and a wide and easily extendable scope of modalities — and is consequently used by numerous analysis and processing tools as the preferred input format in many fields of neuroscience. HeuDiConv (Heuristic DICOM Converter) enables flexible and efficient conversion of spatially reconstructed neuroimaging data from the DICOM format (quasi-ubiquitous in biomedical image acquisition systems, particularly in clinical settings) to BIDS, as well as other file layouts. HeuDiConv provides a multi-stage operator input workflow (discovery, manual tuning, conversion) where manual tuning step is optional and thus the entire conversion can be seamlessly integrated into a data processing pipeline. HeuDiConv is written in Python, and supports the DICOM specification for input

parsing, and the BIDS specification for output construction. The support for these standards is extensive, and HeuDiConv can handle complex organization scenarios such as arise for specific data types (e.g., multi-echo sequences, or single-band reference volumes). In addition to generating valid BIDS outputs, additional support is offered for custom output layouts. This is obtained via a set of built-in fully functional or example heuristics expressed as simple Python functions. Those heuristics could be taken as a template or as a base for developing custom heuristics, thus providing full flexibility and maintaining user accessibility. HeuDiConv further integrates with DataLad (Halchenko et al., 2021), and can automatically prepare hierarchies of DataLad datasets with optional obfuscation of sensitive data and metadata, including obfuscating patient visit timestamps in the git version control system. As a result, given its extensibility, large modality support, and integration with advanced data management technologies, HeuDiConv has become a mainstay in numerous neuroimaging workflows, and constitutes a powerful and highly adaptable tool of potential interest to large swathes of the neuroimaging community.

Statement of Need

Neuroimaging is an empirical research area which relies heavily on efficient data acquisition, harmonization, and processing. Neuroimaging data sourced from medical imaging equipment, and in particular magnetic resonance imaging (MRI) scanners, can be exported in numerous formats, among which DICOM (Digital Imaging and Communications in Medicine) is most prominent. DICOM data are often transmitted to PACS (Picture Archiving and Communication Systems) servers for archiving or further processing. Unlike in clinical settings, where data are interfaced with directly from PACS in the DICOM format, in neuroimaging research, tools typically require data files in the NIfTI (NIfTI Data Format Working Group, 2003--) format which directly stores images as 3D or 4D objects and restricts metadata to the most useful attributes. Tools such as `dcm2niix` (Li et al., 2016) can be used to convert DICOM files into NIfTI files, and can extract metadata fields not covered by the NIfTI header into sidecar `.json` files. However, the scope of such tools is limited, as it does not extend to organizing multiple NIfTI files for different subjects and possibly scanning sessions within a study.

HeuDiConv was created in 2014 to provide flexible tooling so that labs may rapidly and consistently convert collections of DICOM files into collections of NIfTI files in customizable file system hierarchies. As manual file renaming and metadata reorganization is tedious and error prone, automation is preferable, and this is a consistent focus of HeuDiConv.

Since the inception of HeuDiConv in 2014, the BIDS standard (K. J. Gorgolewski et al., 2016) was established. BIDS standard formalizes data file hierarchies and metadata storage in a fashion which, due to its community-driven nature, is both highly optimized and widely understood by analysis tools. Since then, DICOM conversion to NIfTI files contained within a BIDS hierarchy has emerged as the most frequent use-case for HeuDiConv.

Overview of HeuDiConv functionality

HeuDiConv has been developed to implement logic commonly used across labs (grouping DICOMs, extracting metadata, converting individual sequences, populating standard BIDS files, etc.) while allowing individual groups to customize **how** files should be organized and named while driving custom decisions through the conventions and desires of those individual groups. Such decision making is implemented in *HeuDiConv heuristics*, which are implemented as Python modules following some minimalistic specified interfaces documented in HeuDiConv documentation (https://heudiconv.readthedocs.io/en/latest/heuristics.html). HeuDiConv, if instructed to operate in BIDS mode (`--bids` flag) with a heuristic providing base naming instructions, and helpers to organize the files in the hierarchy defined by the BIDS standard. It also ensures files are named according to the BIDS specifications, including complex composite

recordings such as those associated with multi-echo sequences.

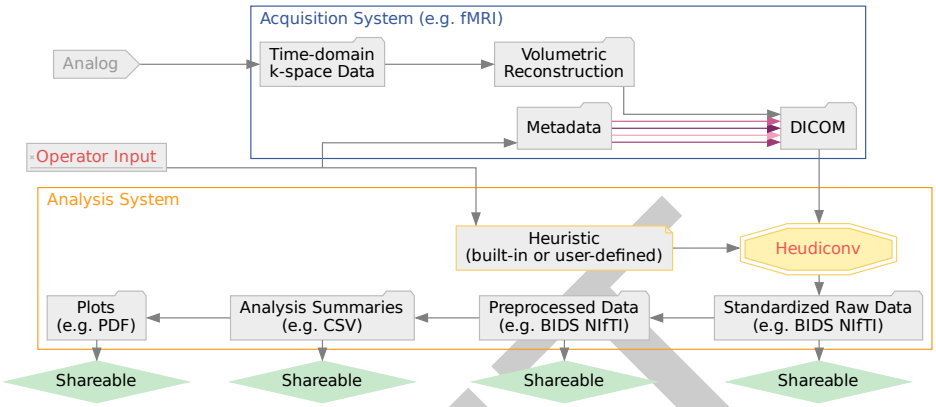

Figure 1: HeuDiConv automates the keystone conversion step in reproducible data handling, without compromising operator flexibility. The showcased set-up depicts a 2-machine infrastructure, with heudiconv operating on the same machine as subsequent analysis steps for data in a standardized and shareable representation. For more advanced usage at institutions with dedicated infrastructure, HeuDiConv can operate on an additional third machine, interfacing between the depicted two, and dedicated to data repositing, versioning, and backup.

Exemplar heuristics

Convertall

The convertall heuristic is the simplest heuristic which expresses no knowledge or assumptions about anything and can be used as a template to develop new heuristics or to establish initial mapping for manual naming of the sequences in the "manual curation" step.

StudyForrest phase 2

The studyforrest_phase2 heuristic is a small sample heuristic developed for the StudyForrest (Hanke et al., 2014) project, and demonstrates custom conversion into BIDS dataset.

ReproIn

The ReproIn heuristic was initially developed at the Dartmouth Brain Imaging Center (DBIC) to automate data conversion into BIDS for any neuroimaging study performed using the center's facilities. The core principle behind ReproIn is the reduction of operator interaction required to obtain BIDS datasets for acquired data. It is achieved by ensuring that reference MRI sequences on the instrumentation are organized and named in a consistent and flexible way, such that upon usage in any experimental protocol they will encode the information required for fully automatic conversion and repositing of the resulting data.

In case of correct specification and absent operator errors, such as mis-typed subject or session IDs, it can be fully automated, and work is ongoing to make such deployments turnkey. Visit ReproIn project page http://reproin.repronim.org to discover more.

Adoption and usage

As a citeable resource RRID:SCR_017427, Heudiconv has already 6 mentions in papers at time of writing. There is a growing number of downloads from PyPI and uses of HeuDiConv (see Figure 2). Over 40 BIDS datasets were converted over to BIDS with HeuDiConv at Dartmouth Brain Imaging Center (DBIC), using the ReproIn heuristic developed there. HeuDiConv was found to be used for PET data conversion (Jamadar et al., 2021), shared as OpenNeuro ds003382 (Sharna Jamadar et al., 2020). Moreover, the HeuDiConv approach inspired the development of `fw-heudiconv` (FlywheelTools: Software for HeuDiConv-Style BIDS Curation On Flywheel) (Tapera et al., 2021).

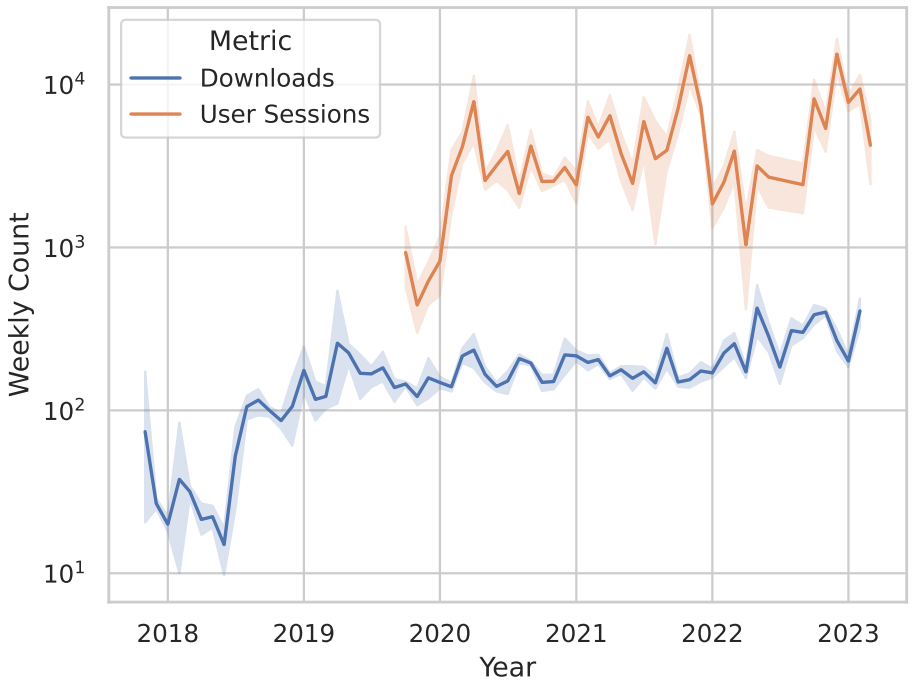

Figure 2: Downloads experienced an initial sharp rise after the ReproNim HeuDiconv training event at OHBM in mid 2018, and have subsequently followed a positive trend along with the usage — exceeding 1000 sessions per week — in the data collection interval. Depicted are weekly download and confirmed session estimates, averaged per month, with a 95% confidence interval. User session estimates for July and August 2022 are linearly extrapolated from the nearest neighbour. Download counts are sourced from PyPI, the Python community repository; whereas user session counts are sourced from Etelemetry, an infrastructure for verifiable research impact, which end-users can disable to protect privacy.

External dependencies

HeuDiConv uses specialized tools and libraries:

- `datalad` (Halchenko et al., 2021) (RRID: SCR_003931) enables managing produced datasets as version controlled repositories.
- `dcm2niix` (Li et al., 2016) is used for the conversion from DICOM to NIfTI and initial versions of sidecar .json files,
- `etelemetry` and `filelock` are used as supplementary utilities,
- `neurodocker` (Kaczmarzyk et al., 2023) (RRID:SCR_017426) is used to produce `Dockerfile` from which docker images are built,

- `nipype` (K. Gorgolewski et al., 2011) (RRID:SCR_002502) to interface `dcm2niix` and extra metadata invocations,
- `pydicom` (Mason et al., 2022) (RRID:SCR_002573) and `dcmstack` for DICOM analysis and extraction of extra metadata to place to BIDS sidecar files,
- `pytest` formalizes unit and integration testing.

Acknowledgments

We would like to extend our gratitude to Matthew Brett, Jörg Stadler, Russell Poldrack, Sin Kim, Dan Lurie, and Henry Braun for notable contributions to the codebase, bug reports, recommendations, and promotion of HeuDiConv.

HeuDiConv development was primarily done under the umbrella of the NIH funded Nipype 1R01EB020740-01, ReproNim 1P41EB019936-01A1 and 2P41EB019936-06A1 (PI: Kennedy).

References

Gorgolewski, K. J., Auer, T., Calhoun, V. D., Craddock, R. C., Das, S., Duff, E. P., Flandin, G., Ghosh, S. S., Glatard, T., Halchenko, Y. O., Handwerker, D. A., Hanke, M., Keator, D., Li, X., Michael, Z., Maumet, C., Nichols, B. N., Nichols, T. E., Pellman, J., … Poldrack, R. A. (2016). The brain imaging data structure, a format for organizing and describing outputs of neuroimaging experiments. *Scientific Data*, *3*, 160044. https://doi.org/10.1038/sdata.2016.44

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Halchenko, Y., Meyer, K., Poldrack, B., Solanky, D., Wagner, A., Gors, J., MacFarlane, D., Pustina, D., Sochat, V., Ghosh, S., Mönch, C., Markiewicz, C., Waite, L., Shlyakhter, I., Vega, A. de la, Hayashi, S., Häusler, C., Poline, J.-B., Kadelka, T., … Hanke, M. (2021). DataLad: Distributed system for joint management of code, data, and their relationship. *Journal of Open Source Software*, *6*(63), 3262. https://doi.org/10.21105/joss.03262

Hanke, M., Baumgartner, F. J., Ibe, P., Kaule, F. R., Pollmann, S., Speck, O., Zinke, W., & Stadler, J. (2014). A high-resolution 7-tesla fMRI dataset from complex natural stimulation with an audio movie. *Scientific Data*, *1*(1). https://doi.org/10.1038/sdata.2014.3

Jamadar, S. D., Zhong, S., Carey, A., McIntyre, R., Ward, P. G. D., Fornito, A., Premaratne, M., Shah, N. J., O'Brien, K., Stäb, D., Chen, Z., & Egan, G. F. (2021). Task-evoked simultaneous FDG-PET and fMRI data for measurement of neural metabolism in the human visual cortex. *Scientific Data*, *8*(1). https://doi.org/10.1038/s41597-021-01042-2

Kaczmarzyk, J., Satrajit Ghosh, Goncalves, M., Bollmann, S., Halchenko, Y., Wighton, P., Gau, R., Notter, M., Markiewicz, C., Jarecka, D., Nielson, D., Cieslak, M., Mitchell, R., Mchlrnwld, Araikes, Close, T., Rokem, A., Klein, A., Torres, G., … Sulantha2006. (2023). *ReproNim/neurodocker: 0.9.4*. Zenodo. https://doi.org/10.5281/ZENODO.1058997

Li, X., Morgan, P. S., Ashburner, J., Smith, J., & Rorden, C. (2016). The first step for neuroimaging data analysis: DICOM to NIfTI conversion. *Journal of Neuroscience Methods*, *264*, 47–56. https://doi.org/10.1016/j.jneumeth.2016.03.001

Mason, D., Scaramallion, Mrbean-Bremen, Rhaxton, Suever, J., Vanessasaurus, Orfanos, D. P., Lemaitre, G., Panchal, A., Rothberg, A., Herrmann, M. D., Massich, J., Kerns, J., Korijn Van Golen, Robitaille, T., Biggs, S., Moloney, Bridge, C., Shun-Shin, M., … Wada, M. (2022). *Pydicom/pydicom: Pydicom 2.3.1*. Zenodo. https://doi.org/10.5281/ZENODO.7319790

NIfTI Data Format Working Group. (2003--). *Nifticlib: IO libraries for the NIfTI-1 data format*. http://niftilib.sourceforge.net

Sharna Jamadar, Shenjun Zhong, Ward, P., Carey, A., McIntyre, R., Fornito, A., Premaratne, M., N Jon Shah, O'Brien, K., Stab, D., Zhaolin Chen, & Egan, G. (2020). *Monash vis-fPET-fMRI*. Openneuro. https://doi.org/10.18112/OPENNEURO.DS003382.V1.0.0

Tapera, T. M., Cieslak, M., Bertolero, M., Adebimpe, A., Aguirre, G. K., Butler, E. R., Cook, P. A., Davila, D., Elliott, M. A., Linguiti, S., Murtha, K., Tackett, W., Detre, J. A., & Satterthwaite, T. D. (2021). FlywheelTools: Data curation and manipulation on the flywheel platform. *Frontiers in Neuroinformatics*, *15*. https://doi.org/10.3389/fninf.2021.678403