The DataLad Handbook: A user-focused and workflow-based addition to standard software documentation

Poster No:

1933

Submission Type:

Abstract Submission

Authors:

Adina Wagner[1], Laura Waite[1], Alexander Waite[1], Niels Reuter[2,2], Benjamin Poldrack[1], Jean-Baptiste Poline[3], Tobias Kadelka[1], Christopher Markiewicz[4], Peter Vavra[5], Lya Paas Oliveros[1,2], Peer Herholz[6], Lisa Mochalski[1,2], Lisa Wiersch[1], Nevena Kraljevic[1,2], Marisa Heckner[1,2], Pattarawat Chormai[7], Yaroslav Halchenko[8], Michael Hanke[1,2]

Institutions:

[1]Institute of Neuroscience and Medicine (INM-7), Forschungszentrum Jülich, Jülich, Germany, [2]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Düsseldorf, Düsseldorf, Germany, [3]McGill University, Montreal, QC, [4]Stanford University, Stanford, NH, [5]Department of Biological Psychology, Otto von Guericke University Magdeburg, Magdeburg, Germany, [6]McGill University, Montréal, Quebec, [7]Max Planck School of Cognition, Leipzig, Germany, [8]Dartmouth College, Hanover, NH

Introduction:

Big data has arrived in neuroscience: There is growing awareness of the role of sample size and replicable results (Button et al., 2013; Turner et al., 2018), a rise of platforms, tools, and standards that aim to facilitate data sharing and management (Wiener et al., 2016), and unprecedented sample sizes (e.g., UKBiobank; Bzdok & Yeo, 2017). Together with increasingly complex data analyses, it requires methodological knowledge, but also skills in data and software management, to put open, reproducible, and scalable neuroimaging research into effect. Researchers, planners (PIs or funders), and trainers in neuroscience alike are challenged by these demands (Fothergill et al., 2019, Grisham et al., 2016) and rely on documentation of their tools. Researchers need accessible educational content to understand and use a tool, planners need high-level, non-technical information in order to make informed yet efficient decisions on whether a tool fulfills their needs, and trainers need reliable, open teaching material. But without a computer science background, many neuroscientists receive no formal training in data management. To be broadly accessible, documentation needs to cultivate confidence in users regardless of their background. However, software documentation can be sub-

optimal for novices: It may be incomplete, narrowly focused on individual commands, or assume existing knowledge novices lack (Segal, 2007; Pawlik et al., 2015), and can thereby discourage potential users or inhibit the adoption of valuable tools.
DataLad (Halchenko, Hanke et al., 2019) is an open source data management multi-tool that facilitates digital workflows of data and science. It is used as an infrastructure component or utility in a growing number of services and tools (e.g., OpenNeuro, CBRAIN platform, HeuDiConv, brainlife.io). Despite being accessible (free, open, with support for all major OSs) and comprehensive, adoption among users is impeded by a lack of workflow-based and user-focused educational materials.

Methods:

A user-driven alternative to scientific software documentation by software developers, "Documentation Crowdsourcing", has been successfully employed by the NumPy project (Oliphant, 2006; Pawlik et al., 2015). Extending this concept beyond documentation, we have created the DataLad handbook (handbook.datalad.org) as a free & open-source, user-driven and -focused educational instrument and resource for trainers, users, and planners for (research) data management, independent of their background and skill level. The handbook contains three components: 'The Basics', a course-like, continuous narrative that integrates basic functions and concepts into larger workflows as a 'code-along' crash course serves as a tutorial. Recipe-like step-by-step how-to-guides ('use cases') show high-level overviews for specific applications. Expandable sections contain background information and serve as explanations. The code is tested automatically to ensure correct operation, and the comprehensibility of contents is tested and improved via handbook-based teaching sessions for data management novices (various career stages, MSc students up to PI-level). Slides and automatic code demos can be semi-automatically extracted as readily available, tested resources for teachers.

Results:

The handbook serves several needs: 1) learning-oriented and lesson-like tutorials allow novices to get started, 2) goal-oriented how-to-guides show planners how to solve specific problems, 3) optional explanations provide background and context beyond what is necessary to use a concept or command, and 4) a continuously tested codebase allows trainers to get reliable teaching materials with reproducible examples.

Conclusions:

We present an accessible, free, and open source educational instrument to explain and demonstrate state-of-the-art data management procedures with DataLad. The framework behind the project is generic, and could be reused for similar training materials.

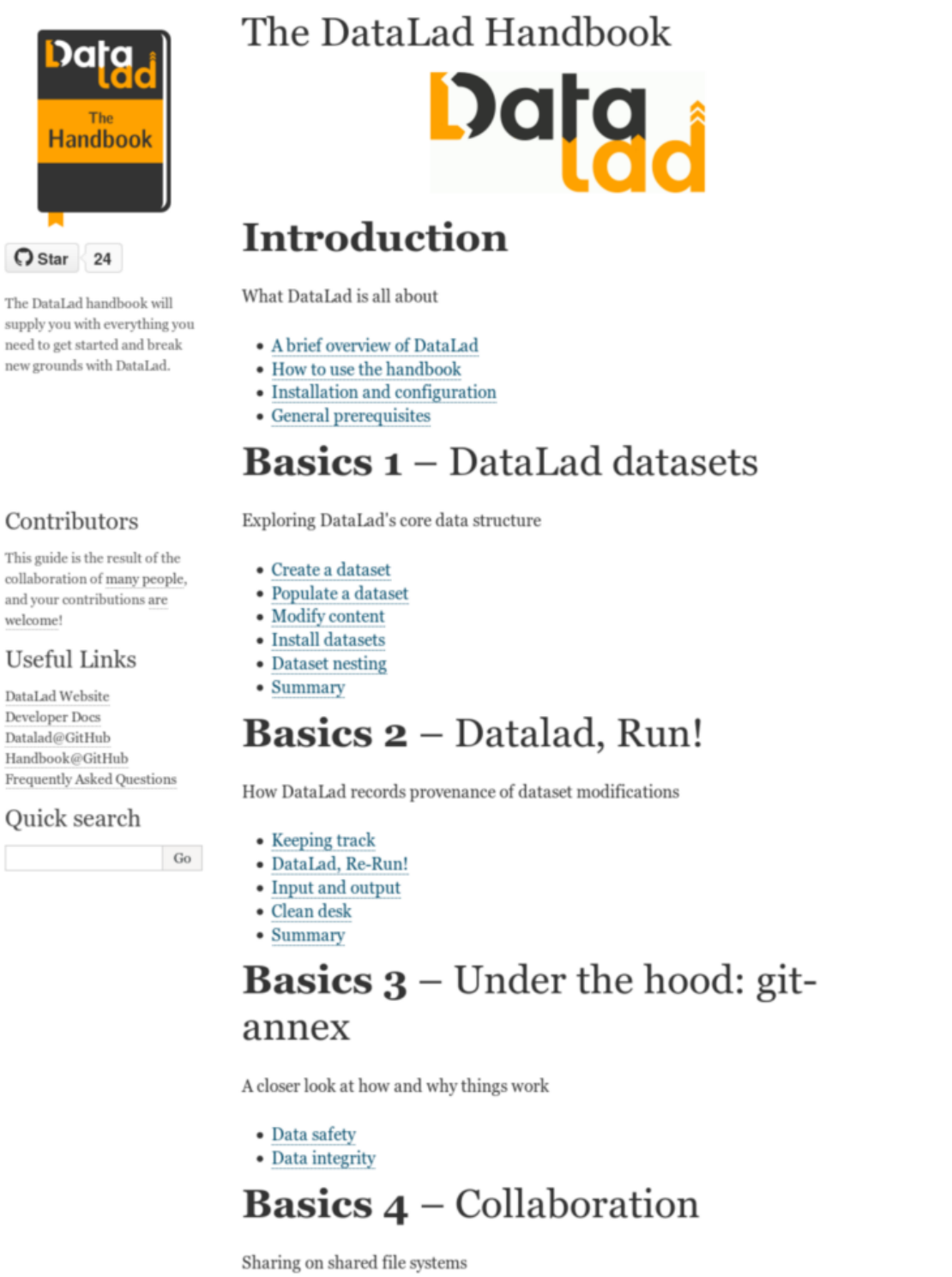

Figure 1: Excerpt of the HTML index of the handbook at handbook.datalad.org.

Figure 2: A photo montage of all pages in the PDF version of the book.

Neuroinformatics and Data Sharing:

Databasing and Data Sharing [1]
Workflows
Informatics Other [2]

Keywords:

Data Organization
Informatics
Other - Research Data Management ; Teaching ; Education

[1|2]Indicates the priority used for review

My abstract is being submitted as a Software Demonstration.

No

Please indicate below if your study was a "resting state" or "task-activation" study.

Other

Healthy subjects only or patients (note that patient studies may also involve healthy subjects):

Healthy subjects

Was any human subjects research approved by the relevant Institutional Review Board or ethics panel? NOTE: Any human subjects studies without IRB approval will be automatically rejected.

Not applicable

Was any animal research approved by the relevant IACUC or other animal research panel? NOTE: Any animal studies without IACUC approval will be automatically rejected.

Not applicable

Please indicate which methods were used in your research:

Other, Please specify - Research data management education

Provide references using author date format

Button, K., et al. (2013) Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14, 365–376 doi:10.1038/nrn3475
Bzdok, D. et al. (2017). Inference in the age of big data: Future perspectives on neuroscience, NeuroImage, 155, 549-564. https://doi.org/10.1016/j.neuroimage.2017.04.061.
Fothergill, B. T., Knight, W., Stahl, B. C., & Ulnicane, I. (2019). Responsible Data Governance of Neuroscience Big Data. Frontiers in Neuroinformatics, 13, 28. doi:10.3389/fninf.2019.00028
Grisham, W., Lom, B., Lanyon, L., & L Ramos, R. (2016). Proposed training to meet challenges of large-scale data in neuroscience. Frontiers in Neuroinformatics, 10, 28.
Halchenko, Y. O., et al. (2019, October 20). datalad/datalad 0.12.0rc6 (Version 0.12.0rc6). Zenodo. http://doi.org/10.5281/zenodo.3512712
Oliphant, T. E. (2006). A guide to NumPy (Vol. 1). Trelgol Publishing USA.
Pawlik, A., et al. (2015). Crowdsourcing Scientific Software Documentation: A Case Study of the NumPy Documentation Project, in Computing in Science & Engineering, 17(1), pp. 28-36. doi: 10.1109/MCSE.2014.93
Segal, J (2007). Some problems of professional end user developers. In: Visual Languages and Human-Centric Computing (Cox, Philip and Hosking, John eds.), IEEE Computer Society/Conference Publishing Services, Los Alamitos, Ca, USA pp. 111–118.
Turner, B.O., et al. Small sample sizes reduce the replicability of task-based fMRI studies. Commun Biol 1, 62 (2018) doi:10.1038/s42003-018-0073-z
Wiener, M., et al. (2016). Enabling an Open Data Ecosystem for the Neurosciences, Neuron, 92(4).