Personal Stories

Print to PDF

Contents

An Interview with Adina on Datalad

Version controlling data can be challenging. Adina knows this because she is part of a team that develops DataLad and uses it to solve data management challenges. Kirstie interviews her about her work and why she thinks versioning data is essential.

Kirstie: Hi Adina, thank you for contributing the chapter on version control for data! I know you are a developer for DataLad, and I'm excited to learn more about the project. Can you start by telling me who you are and what you are working on?

Adina: Hey Kirstie, thanks a lot for providing a space for the topic of version-controlling data! I'm a PhD student in neuroscience, and I am part of the lab that develops DataLad. Apart from working on neuroscientific questions, I also work on data management challenges that are typical for my field, such as "I have 300GB of data, how can I possibly version control or share this?", or "How can I link my analyses to the version of data I have used?". As a neuroscientist, I'm privileged to work in a field with many fantastic, open data sets, but it is also challenging to handle, share, and keep track of data that can easily be several hundred GB in size.

Kirstie: Fab, so how does DataLad help with your work?

Adina: DataLad lets me version control and share data of any size, and I use this to attach data in precise versions to code and manuscripts I create. When doing data analyses and the underlying data is modified, I can update my repositories and recompute my scripts. This helps me to assess if my results are replicable. And just as Git, it is a great memory aid for remembering what I did to my data. It has some cool functions for provenance capture, and I can just check my Git history to find out from which data a particular figure was created, for example.

Kirstie: Cool, so what makes DataLad better suited for what you do than other tools that version control data?

Adina: I personally like DataLad, because on top of the functionality that Git and `git-annex` provides, it makes linking and reusing modular parts of my research easy. When I work on an analysis, I publish the data, the code + results, and the manuscript as separate, version-controlled Git repositories to GitHub. But these repositories are linked together so that someone who reads my manuscript could backtrace every step that was undertaken to create this result, back to the original data. I can share my analysis on GitHub and can have data, code, and even software environments altogether, to allow others to reproduce my results, and I find that to be a very powerful feature.

Kirstie: And as a part of the DataLad team, how do you contribute to the software?

Adina: My main motivation is to make the software accessible for users of all backgrounds. If scientists receive no formal training in version control or research data management, it can be hard to work reproducibly. I believe if software is easy to use and well-documented, it can help scientists to do better science. Software-wise, I, therefore, work on help- and UX-features, and documentation-wise, I work on tutorials that are suitable to users independent of skill level or background.

Kirstie: What is the journey of DataLad, and how did you get to be a part of it?

Adina: DataLad was originally created by Michael Hanke and Yarik Halchenko in 2014. They wanted to have a tool that allowed them to install data just as easily as software packages and keep track of how data changes. `git-annex` already existed at this point, but they wanted to build upon it to make it easier to use. Over the years, the tool became a joint version control and data management tool to facilitate data sharing, revision tracking, and reproducible computations. I joined the lab almost two years ago as a Master's student in Clinical Psychology, excited for open and reproducible science, but a complete newbie technology-wise: I had never heard of version control, no programming experience, and the idea that data is dynamic was insightful but completely new to me. Naturally, when I started using DataLad, I was completely overwhelmed. Luckily, there were many people to help me get started and give me the necessary background information. I know, however, that such a learning environment is not the default, so when I started my PhD, I actually created the resource that I would have needed to get started as a student: The DataLad Handbook.

Kirstie: Thanks a lot for telling us about this tool. So the handbook is where people can find out more, if they want?

Adina: Yes, I would point them to The DataLad Handbook. It is meant to be an accessible, code-along tutorial, that is suitable to researchers independent of background - I think you

shouldn’t have to be a Linux-crank or computer scientist to version control data.