**Data Version Control**

This year, I’ve been selected as a Fellow Freies Wissen to further our understanding of data version control.

**What is data version control?**

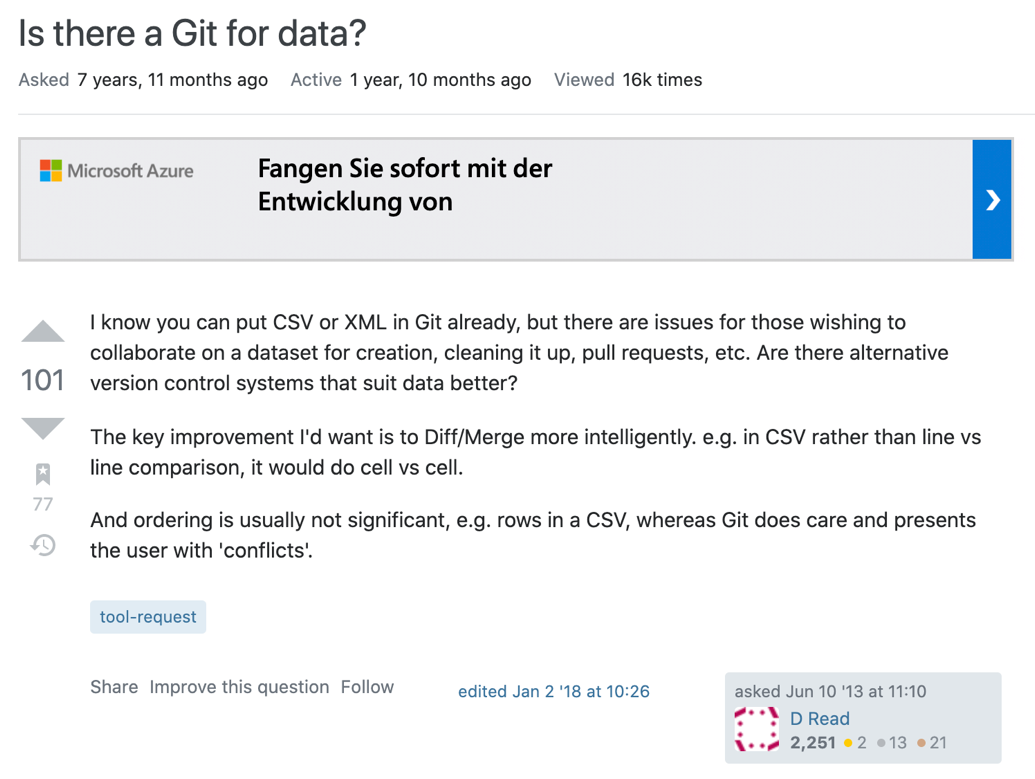
Real world data, such as logs, comments, and reviews, are usually messy and require substantial processing. Additionally, this data may be updated during the project, e.g., when data collection is ongoing (as is the case for many COVID-19 working papers). This can result in many different versions of datasets throughout the research project (e.g., v2, v3, featureXpresent, featureYabsent) and may lead to confusing and hardly reproducible results when looking at an analysis that has been conducted a few weeks back. Moreover, the data that researchers in social sciences and other fields have at hand become increasingly large, often exceeding several GBs. Simply running the entire pre-processing pipeline when compiling an R Markdown document can take hours to complete. Versioning intermediate datasets with Git is often not feasible due to storage restrictions of code repositories and technical difficulties to efficiently handle large files.

**Are researchers already incorporating data version control?**

As part of my fellowship, I launched a survey to understand research practices among other fellows and alumnus. In particular, I asked them how they were versioning the different parts of their research endeavor, such as the manuscript the data and the scripts. Although version control was very prevalent for manuscripts and data analysis scripts, version control of the data was not so common. This clearly shows a gap in the current pipeline.

**Which tools exist?**

“Git for data” has been a question originally posted on Stack Overflow.



Since then many tools have been developed to deal with version control of data. One list can be found here: <https://docs.google.com/spreadsheets/d/1jGQY_wjj7dYVne6toyzmU7Ni0tfm-fUEmdh7Nw_ZH0k/edit?ts=5fc6a2a5#gid=0>

Git for data specifically allows for the versioning of different data sets and pipelines. As researchers have increasingly embraced git to version control manuscripts and scripts, version control of data is not on the radar of many researchers.

<https://www.youtube.com/watch?v=r5uxntl_hWg>

**Which tools should researchers use?**

It depends on the use case. As outlined in the video shown above, there are different kinds of tools available, some offering more of a database-like tool where you can store data and see version control of that data, others offer you a more replicable research pipeline. Many (not all) are open source, which would definitely be an advantage.

Dolthub, for example, are advertising their space as a version controlled database which different people can make additions.

DVC is a pipeline tool that’s lightweight and storage agnostic.