This is the “decade of big data” [2]. In the field of sequence analysis, for example, datasets the size of several gigabytes are frequently encountered. The availability of big datasets, though, brings about new challenges not only for the analysis methods but also for the handling of the data – especially in connection reproducible research [4, 3].

Since reproducibility demands the data to be present in a specified version, version control systems for both code and data are of great help to the researcher. Until recently, the available version control systems failed to handle big data well. As a result, massive datasets are seldom put under version control and, thus, are often kept separate from the project.

The novel git-annex [1] extension to the git version control system closes this gap. As the original target of git-annex were multimedia files, this system is tailored with a special focus on big data files.

We present a wrapper from R to git-annex that makes it easy to handle big datasets from within the analysis project. With this method, also big data finally becomes part of the project repository and is easily accessible. In addition, the content of the large files can be safely moved away from the working directory and backups get automatized.

References


