Geophysical Research Abstracts
Vol. 20, EGU2018-4617, 2018
EGU General Assembly 2018

Scientific Reference Datasets - Cornerstones of Reproducible Research: Design, Test & Distribution using R

Sebastian Kreutzer (1), Christoph Burow (2), Mathieu Duval (3), and Geoff Duller (4)

(1) Université Bordeaux Montaigne, IRAMAT-CRP2A, Pessac Cedex, France (sebastian.kreutzer@u-bordeaux-montaigne.fr), (2) Institute for Geography, University of Cologne, Germany, (3) Australian Research Centre for Human Evolution, Environmental Futures Research Institute, Griffith University, Australia, (4) Department of Geography and Earth Sciences, Aberystwyth University, United Kingdom

Modern science demands more and more custom-tailored software solutions. Researchers usually respond to this demand by developing numerous programmes by themselves. Either to bridge the gaps between commercial solutions or to provide workflow enhancements. Supported by a variety of freely available tools, the development of scientific software packages was never more comfortable and its distribution via public repositories is only a few clicks away. Users are being offered multiple software solutions to tackle a particular problem. While independently developed programmes can appear comparable regarding features, they may not return similar results. However, understanding differences between software solutions proves difficult. In complex analytical pathways finding the origin of diverging results quickly becomes a time-consuming task, regardless of the questions whether the source code is available or not. That is where scientific reference datasets come into play; based on real measurements or artificially generated. For developers, working with reference datasets ensures accurate implementation at the development stage of the analytical processes without scrutinising other people's code. For users, reference datasets are easy means to test the reliability and reproducibility of the software.

We present and discuss rationale, design and limitation of reference datasets, exemplified for the field of luminescence and electron spin resonance (ESR) dating. For the creation, distribution and versioning of the reference dataset, we use the statistical programming language **R** [1]. **R** has become the preferred all-purpose coding solution for many scientists and gained growing attention in the field of luminescence and ESR dating [2]. Together with the Comprehensive R Archive Network (CRAN) the **R** environment provides a robust and transparent platform to generate, document and distribute reference datasets as a separate package. Moreover, reference data stored within an **R** package are not limited to test software developed with **R**. For the example of luminescence and ESR data, we show how bundle data with further **R** functions. Reference data can be created on demand and exported into free or proprietary formats. Our contribution intends to stimulate discussions on the question to which extent the concept of reference data generation and distribution using the **R** environment can be transferred to other scientific fields to improve the reliability and reproducibility of scientific software.

References

[1] R Core Team, 2018. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://r-project.org

[2] Kreutzer, S., Burow, C., Dietze, M., Fuchs, M.C., Fischer, M., Schmidt, C. 2017. Software in the context of luminescence dating: status, concepts and suggestions exemplified by the R package 'Luminescence'. Ancient TL 35, 1–11. URL: http://www.ecu.edu/cs-cas/physics/Ancient-Timeline/upload/ATL_35-2_Kreutzer_p1-11.pdf