February 10th, 2020

Template for reproducible scientific datasets/papers

Mohammad Akhlaghi
Big Data Astronomer at the Instituto de Astrofísica de Canarias (IAC), Spain

Abstract

The increasing volume, diversity, and role of data in modern research has been very fruitful. However, these same factors, have also made it harder to describe (in sufficient detail) the processing behind a scientific result within the confines of a traditional paper. It is thus becoming harder and harder to reproduce (i.e., critically review by coauthors, referees or larger community) results that define scientific progress. In this talk, a working solution to this problem is proposed.

It is (a template) [https://gitlab.com/makhlaghi/reproducible-paper/blob/master/README-hacking.md] that provides a framework to exactly reproduce a scientific analysis (from the input data and software, to the processing and creation of final report/paper/dataset. The necessary software are built (from the low-level C compiler and shell, to the higher-level science programs and all their dependencies) with the predefined configuration. The software are then run on the input data sets to produce the final result.

The template will finally produce a “dynamic” PDF using LaTeX macros: any change in the analysis will automatically update the relevant parts of the PDF (for example numbers, tables or figures). A project defined in this template is fully managed and published in plain text and only consumes a few hundred kilo-bytes (unlike binary blobs like Docker, although building it in a Docker image is trivial). It is thus easily to publish (for example on arXiv with the paper’s LaTeX source), and give readers the ability to exactly reproduce the paper’s results if they need. It is also easily search-able, providing a treasure trove to extract metadata on the project (even after publication, and without the author’s active involvement). This can be very valuable when implement widely (e.g., using machine learning on many project sources to define automatic workflows). But most importantly it will allow other scientists to independently study the details, verify in practice, and build incrementally on each others’ work, without necessarily needing to run it.

Short bio

Mohammad Akhlaghi is a Big Data Astronomer at the Instituto de Astrofísica de Canarias (IAC), Spain. He is the founder of a (reproducible paper template) [https://www.rd-alliance.org/node/64603] project that was awarded an RDA-Europe adoption grant, and also the founder of (GNU Astronomy Utilities) [https://www.gnu.org/software/gnuastro] (a collection of free software programs and libraries for data analysis).

He received his PhD in astronomy from Tohoku University (Japan) and prior to coming to the IAC, he was CNRS postdoctoral fellow in Centre de Recherche Astrophysique de Lyon (CRAL, France).