FAIR²

Claudio D'Onofrio[1], Karolina Pantazatou[1], Ute Karstens[1], Margareta Hellström1, Ida Storm[1], and Alex Vermeulen[2]

Reuse – Reproduce – Replicate : Long-term Reproducibility for Jupyter Notebook

Status Quo

Jupyter Notebooks are a popular choice for interactive scientific computing to convey descriptive information together with source code. Such 'executable' documents provide a paradigm shift in scientific writing, where not only the science is described, but the actual computation is open available and with the aim of being reusable and reproducible, hence can be independently validated.

Reuse & Reproduce : Longevity

The reality is, that most public available Jupyter Notebooks do not run. Pimentel et al. analysed over 800'000 Jupyter notebooks from GitHub. 24 % executed without errors and only 4 % produced the same results. We suspect that the combination of fast development cycles in open source software and CI/CD is one of the main reasons responsible for a low shelf life of published Jupyter Notebooks. Even if the computational environment is fully described, over time it isalmost impossible to re-create the user's installation. Further it is often neglected to fully describe all the dependencies (if they are known) to the user.

The information provided with the FAIR principles are a necessity, no arguments there, but most likely you need expert knowledge to replicate the method or computation. ReScience (Rougier et al.), showed, that the real added value would be to publish a paper with an independent replication of the scientific work. Than the knowledge is preserved which automatically leads to longevity.

Remove Complexity

We propose to archive the docker image, the user space (user installed packages and settings) and finally the source code. Recreating the system in this way is more like restoring a backup, where backup is the equivalent of an entire computer system. It does not solve all the problems but removes a great deal of complexity and uncertainty.

Replicate

Creating a 'snapshot' of the user environment, allows to persistent identification for a document is technically easy. The FAIR principles are a very good guideline to describe the content of the document with all the nitty gritty details. However the real added value comes from replicating the method or results described in the published article to validate and preserve the knowledge, not the computational environment. But as a first step, the experiment must be reproducible.

Caveats & Limitations

For our application we do allow by default to access external resources. Which in return means, the longevity is limited to the availability and consistency of these resources. Ideally you refer only to persistent identified digital objects. Our 'snapshot' includes the messy-ness of users, but allows to understand what is happening for a real replication.

References

Rougier et al, 2017, Sustainable computational science: the ReScience initiative. https://doi.org/10.7717/peerj-cs.142
Pimentel et al. A Large-Scale Study About Quality and Reproducibility of Jupyter Notebooks. https://doi.org/10.1109/MSR.2019.00077
Wilkinson, The FAIR Guiding Principles for scientific data management and stewardship. https://doi.org/10.1038/sdata.2016.18

Message to take home

FAIR principles are necessary and very good, but not enough. We propose to emphasize **reproducible** for longevity of scientific computational notebooks as a step to add real value to the science and preserve the knowledge. The scientific community needs to be able to **replicate** the method, experiment, result.

Affiliations: **1** ICOS Carbon Portal, Lund University, Lund, Sweden (claudio.donofrio@nateko.lu.se) **2** ICOS ERIC - Carbon Portal, Lund University, Lund, Sweden