Big data, big responsibility: recording a project's data lineage for publishing reproducible results

RESEARCH DATA ALLIANCE

INSTITUTO DE ASTROFISICA DE CANARIAS · IAC ·

The Instituto de Astrofisica de Canarias (IAC) has designed a "reproducible paper" template by following best practices, including RDA recommendations and outputs, in their research since 2015. It has significantly grown since then, and is now a fully documented template. With the support of an RDA Europe 4.0 grant, the project team, led by Dr Mohammad Akhlaghi, is aiming to improve, test, and promote our adoption of RDA guidelines, and in particular the "Workflows for Research Data Publishing" recommendation and output.

The challenge

Modern data analysis involves a very large and complex series of steps: including the various sources of input data (from different databases), to the software used (and their precise versions, installation configurations and host operating system), and finally the recipes of how the software are run on the data (and their ordering) to produce the final results (added-value datasets, or the figures, plots and tables in a published scientific paper). A project's data lineage contains all this information, as well as their links (for example, which analysis step should be done after which). Without the lineage, the results can't be precisely reproduced/validated by others (or even by the same team!). However, the traditional format of scientific papers is not friendly when it comes to publishing the full data lineage (otherwise each paper would be +50 pages!). Buckheit & Donoho (https://doi.org/10.1007/978-1-4612-2544-7_5) summarize the situation very nicely: "An article about computational science is not the scholarship itself, it is merely advertising of the scholarship. The actual scholarship is the complete software development environment and the complete set of instructions which generated the figures." Our challenge was thus to design a framework that would allow easy recording, preservation and publication of a project's full data lineage.

The RDA outputs adopted

We have adopted the guidelines of the shared RDA and World Data System (WDS) working group on "Workflows for Research Data Publishing: Models and Key Components". By formalizing key components in publishing of data and providing a reference model for it, this output provided a very good set of pointers for us to better formalize our solution. Of course, many other RDA Working Groups and forum discussions also greatly inspired us in the five years that Maneage was brewing in the background of our astronomical data analysis challenges.

Benefits of adoption and impact

Prior to the publication of a project as a scientific paper, adopting Maneage has the following advantages:

1) The full history of the various analysis steps and software versions is recorded.

2) Changing any step (to see its effect on the final result) is trivial because all the numbers in the text, as well as plots and figures, are automatically generated.

Find out more at:

www.rd-alliance.org/recommendations-outputs

Visit rd-alliance.org

or write us at enquiries@rd-alliance.org

3) Geographically separated co-authors can exactly reproduce an on-going project on their independent computers, contribute their analysis steps to a single project in a single work-flow, and later merge it with the work of other team members.

4) It is more straightforward to revert back to a previous state of the project and test an alternative analysis method (to "merge" if it is good).

After the project's publication, readers can exactly reproduce the PDF paper and accompanying data behind the plots. They can easily change any configuration parameter (without necessarily understanding the low-level implementation detail) and see its effect on the plots, tables or numerical results. If they find an interesting result, they can simply write a short paper, only describing the respective change and its consequences, instead of the current situation where authors have to spend the majority of the paper showing that their independent implementation is similar to a previous work. Also, several years after publication, the project can be "merged" back into the improved Maneage "branch" to be usable on systems that weren't available when the paper was originally written.

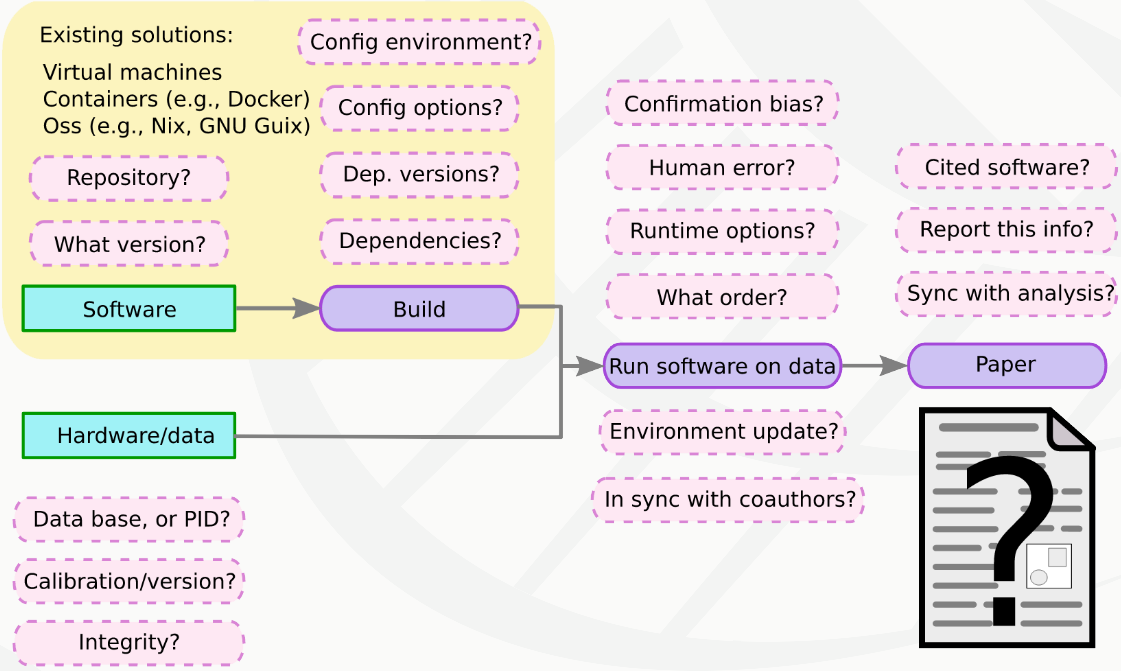
The adoption process

The experimentation with implementation details of what finally became Maneage started about one year before I learnt about RDA (in 2015 with Akhlaghi & Ichikawa, https://doi.org/10.1088/0067-0049/220/1/1). In that project, I was able to formalize and automate the linkage of each analysis command with the published text, and was also able to publish the full series of analysis commands, which can be done automatically, with the paper's LaTeX source on the pre-print service, arXiv: https://arxiv.org/abs/1505.01664. In the next year, I became familiar with the RDA through Francoise Genova's talk in the 26th Astronomical Data Analysis Software & Systems conference in Trieste, Italy. I joined the RDA immediately afterwards and the discussions in the Working Groups that I joined (including Workflows for Research Data Publishing) heavily inspired and influenced me. In my subsequent research projects, I was able to use what I had learnt thanks to the RDA and was able to separate the core infrastructure into a template format, to easily customize for my next individual projects, detailed here https://doi.org/10.5281/zenodo.1163746 and https://doi.org/10.5281/zenodo.1164774. In my subsequent projects, the core template grew considerably and before the 2019 RDA Europe adoption grant applications, it was already being used by R. Infante-Sainz and together we also added the steps to automatize the analysis software installation (which was very complex).

The RDA Adoption grant of 2019 greatly boosted the importance of our meta-project (project on how to do other projects!) within our immediate community (at the IAC), and larger community, where we presented it in seminars. Thanks to the grant, improving and disseminating Maneage came to the forefront of my activities and we were able to invite several early career researchers to the IAC to help in "Maneaging" their scientific projects. The basic design principles of Maneage have also been submitted to a journal (available in arXiv: https://arxiv.org/abs/2006.03018) and we also set up a web presence at https://maneage.org. Recently, Maneage was also used in a COVID-19-related study completely unconnected to astronomy (details here: https://arxiv.org/abs/2007.11779), and several other research teams have also started adopting it. All of our projects (involving the data reduction of several IAC telescopes) are now fully Maneage-based and more teams are adopting it, we will be continuing to develop and improve Maneage. We have also had contacts with IACTec (Technology transfer division of IAC) to help in adopting Maneage in industrial contexts also.

Lessons learned

The Adoption grant initiative in the RDA Europe 4.0 was instrumental for Maneage, in particular due to the fact that we are astronomers and this meta-project is unfortunately not officially considered to be "astronomy" research! With the grant, we were able to convince our institute that investing more time and energy in it is academically productive. It would thus be great if RDA could support other scientists adopting the RDA outputs in their scientific papers. Currently most 2019 adoption grant recipients were data center curators. The reason is that following best practices in data management is unfortunately considered an unnecessary distraction by many scientists, who simply judge a team's outputs by number of published papers. Academically recognized grants like this can help scientists in any field focus on improving the methods of their research and thus gradually improve the culture of data management in the larger scientific community that they belong to (which is a critical component of any science field in the 21st century). This can be accompanied by RDA-sponsored workshops in science institutes on lower-level data curation and management implementation methods.



Existing solutions:

Virtual machines
Containers (e.g., Docker)
Oss (e.g., Nix, GNU Guix)

Config environment?
Config options?
Dep. versions?
Dependencies?
Repository?
What version?
Software
Build
Hardware/data
Data base, or PID?
Calibration/version?
Integrity?
Confirmation bias?
Human error?
Runtime options?
What order?
Run software on data
Environment update?
In sync with coauthors?
Cited software?
Report this info?
Sync with analysis?
Paper

Source: maneage.org

About IAC

The Instituto de Astrofísica de Canarias (IAC) is a public research consortium that is a center of reference within the Spanish astrophysics community, but also in European and worldwide contexts. The IAC maintains the Teide Astronomical Observatory in Tenerife and the Roque de los Muchachos Observatory in La Palma, which is one of the few places on earth with such a high ratio of telescopes to data creators. IAC is also a partner in many international facilities, including the large astronomical projects Euclid and LSST (which will create peta-byte scale datasets of the sky during this decade). The 'Maneage' framework which was the focus of the RDA Europe 4.0 grant was developed to deal with the complex problem of reproducibility in the results of projects that use large and complex sets of datasets. In addition, it has also been designed in a modular and generic framework that is applicable to any data-intensive research project.

Maneage
Managing Data Lineage

Contacts:
Mohammad Akhlaghi