DOI: 10.18727/0722-6691/5281

Report on the ESO Workshop

Reproducibility and Open Science in Astronomy

held online, 10–12 May 2022

Sascha Zeegers[1]
Peter Scicluna[2]

[1] Institute of Astronomy and Astrophysics, Academia Sinica, Taiwan
[2] ESO

Reproducible and open research lies at the heart of science, for both practical and philosophical reasons. To validate results, researchers must be able to access the data and software used to produce them. Meanwhile, as a public good, the outcomes of (publicly-funded) research should be freely available. These were the main topics of the ESO workshop Reproducibility and Open Science in Astronomy, which was held online on 10–12 May 2022, hosted in Santiago, Chile. The goal of the workshop was to discuss the relevance of reproducible workflows in astronomy and potential pathways for the astronomical community. During the workshop the speakers and participants shared examples of reproducible work as well as tools and techniques for improving reproducibility and for mining astronomical data. All talks, tutorials and discussion sessions were recorded and can be viewed online.

The way we do science is constantly evolving and is currently undergoing a paradigm shift, partly driven by the general tendency towards openness in society and the idea that the outcomes of publicly-funded science should be freely available. The ever-increasing volume and complexity of astronomical data and simulations call for accessible methods to verify and reproduce results. Existing and upcoming large facilities (for example, the Atacama Large Millimeter/submillimeter Array, the Square Kilometre Array, the Sloan Digital Sky Survey 5, Vera C. Rubin Observatory, ESO's Extremely Large Telescope and other in this category) provide or will provide us with enormous volumes of data, promising revolutionary scientific capabilities. To handle these large datasets, we can no longer rely on small computing facilities and will need to move analyses to science platforms. It therefore makes sense that now is the time to adapt our approaches to science. Large collaborations aren't alone in benefitting from open science and reproducible research; any individual researcher might struggle to recall details of a previous project. Incorporating open and reproducible workflows at the outset of research could help promote efficient retrofitting and new developments.

The metrics currently used to reward 'excellence' in astronomy — and science in general — do not always benefit open science. Publishing many papers that focus on results and not methods is not beneficial to astronomy in the long term. Promoting alternative metrics to evaluate candidates during hiring processes is important if we are to shift the community's focus.

Aims of the workshop

The workshop Reproducibility and Open Science in Astronomy (ROSA 2022) aimed to share examples of reproducible work as well as tools and techniques for improving reproducibility or mining astronomical data. The experts, brought together by the scientific organising committee (SOC), focused on the questions:

1) What are reproducibility and open science?
2) How to make research reproducible?
3) What are the objectives, benefits, and difficulties of reproducibility?

We designed the workshop programme[1] to be as interactive as possible, achieved by organising two one-hour discussion sessions and leaving ample time after each talk for questions. The workshop included tutorials about data platforms, such as ESA datalabs[2], the Spanish Virtual Observatory[3] (SVO), the Canadian Advanced Network for Astronomy Research[4] (CANFAR) and the Square Kilometre Array Observatory[5] (SKAO). Discussions focused on how to share data and the future of open science. To avoid screen fatigue and to cater to multiple time zones the workshop was restricted to five hours each day. All the talks were recorded, allowing participants to view them at their own pace.

The SOC aimed to reach a wide and varied audience. Since ROSA 2022 was an online workshop, there were no travel costs for participants. The workshop itself was free of charge for the participants, enabling us to reach beyond those already acquainted with the topic and in particular allowing a large number of students to attend.

Summaries of talks and highlights from sessions

Introduction to the workshop

Lourdes Verdes Montenegro gave the introductory talk, getting us all up to speed with an overview of current challenges in reproducible research and possible solutions to them. She focused on the current problems with metrics, introducing alternative metrics (for example, the Declaration on Research Assessment, DORA[6]) and the challenges of big data.

Scientific publishing

The second talk, by Chris Lintott, focused on scientific publishing and how a system of paywalls and publishing fees impedes open science. Open Access is a good development, but if its costs fall exclusively on researchers this creates anger, especially when journals make large profits. One solution is to demand a revolution within science or to publish in open journals such as The Open Journal of Astrophysics. However, Chris pointed out that, in contrast to other fields, astronomy journals are community-owned and all of the same quality and standing, so there are not the same commercial pressures or degree of competitiveness. This means that it is easier for astronomy journals to amend their policies without fear of negative impacts; many of the well-known astronomy journals are now changing their policies on citing software and storage of data.

Chris Erdman gave us a roadmap for sharing models, software, and code. Chris emphasised the benefits of publishing and indexing your software. You get a persistent copy of your software while improving the reproducibility, discoverability and awareness of your research. Chris also gave the very valuable tip to take your entire team to workshops, so everyone is on the same page.

Alice Allen, Editor of the Astrophysics Source Code Library[7] (ASCL) elaborated on the topic of citing and indexing software. The ASCL registers code used in articles, provided that the source code is freely available for download. They also carry out research projects and analyse what happens to the software over time by testing weblinks (Allen, 2021). For instance, 11% of the software published in 2015 was unavailable; when the same links were tested in 2021 the number of broken links had increased to 20%. On the bright side, Alice told us that thanks to many recent changes in the community resources and broad efforts across different disciplines, it has become much easier to index and track software.

Jelle de Plaa told us about the steps taken at the Netherlands Institute for Space Research (SRON) to make research within the institute more reproducible. At SRON they formulated a data management plan and formed a team of data stewards to assist the researchers. The team created reproduction packages and regularly organises workshops within the institute. Jelle provided templates of the reproduction packages used at SRON[8].

Big data science

Mohammad Aklaghli's talk "Big data, big responsibility" focused on the long-term preservation of data and code. To reproduce a research project one needs to know details about the software, for example versions and the configuration environment of the code. Compute containers (for example, Docker[9] and Singularity[10]) may seem like good solutions, but these facilities may not be supported forever. These binary files are expensive to archive because of their size and, because they are binary files, they are not searchable without installing them first. Mohammad presented Maneage[11], a framework which solves this problem using simple plain text files.

Many large projects find that they can no longer bring the data to the researcher because datasets are simply too large; instead, researchers need to go to the data. Science platforms providing all the required analysis tools can be the solution.

The workshop programme included a number of talks and tutorials on data platforms.

- Alex Clarke showed us how regional data centres will provide SKAO data to users. The SKAO runs data challenges (CDAs) aiming to familiarise the scientific community with the data products and tools to extract results from them (Bonaldi et al., 2021). The SKAO have organised two challenges so far and Javier Moldon gave further details of CDA2. To encourage reproducibility, submissions received awards if they adhered to reproducible practices. Javier concluded that extensive community training is required to develop familiarity with reproducibility software.
- Vicente Navarro and Marcos Lopez-Caniego gave us a tutorial on the ESA datalabs science platform (Arviset et al., 2021). The platform has been running a closed beta since March.
- Toby Brown presented ARCADE (Major et al., 2019), an Interactive Science Platform in CANFAR and gave us a live demo. Stephen Gwyn gave a tutorial on creating tables in CANFAR with the Your Catalogues (YouCat) facility.
- Stefania Amodeo introduced the European Science Cluster of Astronomy and Particle physics ESFRI research infrastructures (ESCAPE[12]) project. It provides open access long term data usability for astronomy and particle physics with the aim of bringing different communities together. Stefania focused on Virtual Observatory data and gave a demonstration of MOCpy, a Python library allowing the easy creation, parsing and manipulation of MOCs (Multi-Order Coverage maps).
- Guiseppina Fabbiano talked about the virtual observatory, a multi-wavelength, multimessenger, digital sky that can be searched, visualised and analysed. She showed that the data available in the virtual observatory often get reused in different research projects, which shows how important it is to make data publicly available.
- In a series of tutorials by Fran Jiménez-Esteban, we were introduced to the data platform of the SVO.
- Magda Arnaboldi gave a talk about the ESO Phase 3 archive[13]. She discussed how the archive can ensure the legacy value of science data products.
- Carlo Manara showed the advantages of making data from the ODYSSEUS[14] collaboration publicly available. This resulted in many additional projects on top of the original science plan.
- Jakob Nordin presented the software platform AMPEL (Nordin et al., 2019), which facilitates reproducibility in transient astronomy. The software platform fully enables Code-to-Data for the time domain.

Discussion sessions

Aside from the talks and tutorials, we reserved two hours for discussions about sharing data and the future of open science. All the talks and tutorials and a summary of the discussion sessions are available on the workshop's YouTube channel[15].

Main conclusions from the workshop & ways forward

After an intensive three days, we all concluded we had learned a lot about reproducible and open research. The participants came up with excellent recommendations and lots of ideas for improvement. These ideas were focused not only on individual researchers but also on how we can convince the institutes where we work and the journals in which we publish to make necessary changes.

During the workshop we saw some great examples of how science is becoming more open and reproducible on almost every scale and how astronomy is playing a pioneering role. Journals are making an effort to include datasets and software with published papers, large observatories provide understandable pipelines, universities and institutes are developing policies on how their data and work are stored and researchers are providing their codes and data.

However, there is still room for improvement. Many research papers are not yet reproducible and the projects that are reproducible today may not be a few years from now, as methods may become obsolete. We hope ROSA 2022 will encourage more reproducible research projects and we invite everyone to (re-)watch the talks and tutorials.

Demographics

The SOC sought fair representation from the community. We paid attention to all aspects of the workshop, such as the process of inviting speakers and selecting contributed talks. The SOC agreed that the emphasis of the workshop would lie on the invited talks and tutorials. However, the programme did include five slots for contributed talks. We informed anyone applying to give a contributed talk that the space in our programme would be limited, but that they were welcome to submit abstracts. We were happy to receive seven abstracts from which we selected five for the workshop. The abstracts were ranked in quality by the SOC, who only considered the text of the abstracts to avoid biases from other information.

There were 109 participants. Attendees came from six continents (all but Antarctica), with the following percentages:

- Latin America: 17%
- Asia: 24%
- North America: 8%
- Europe: 48%
- Africa: 2%
- Australia: 2%

With 31% students, 15% postdocs, 34% staff and 23% other participants, we had a good representation of the community. Of our speakers, 27% were female and 73% male. This ratio seems to be reflected in the participants among which 29% were female, 2% non binary, 60% male and 9% left the answer blank. We note that of all seven contributed abstracts we received, six were submitted by male speakers and one by a female speaker. In the case of the invited talks we made an effort to obtain the best possible gender balance. However, there is room for improvement concerning the gender balance of the tutorials. Both the ‘Local’ Organising Committee (LOC) and the SOC were spread all over the world from Taiwan to Chile and we organised the event over several time zones.

Acknowledgements

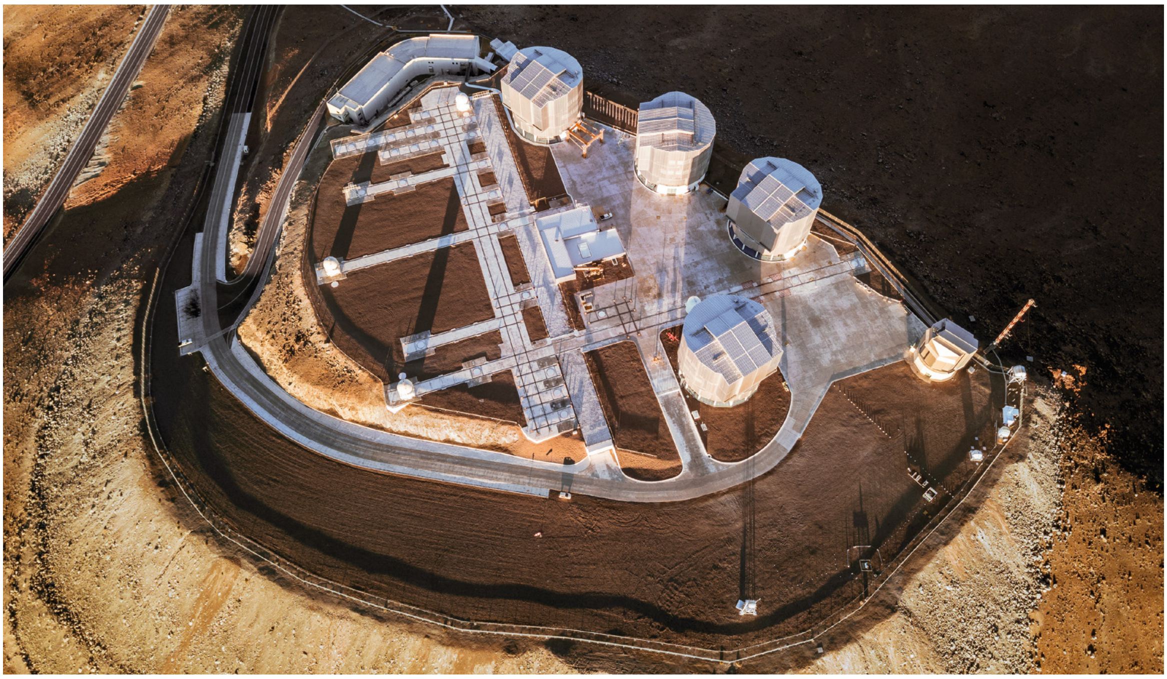
We thank all the participants in ROSA 2022, and especially the speakers, LOC and SOC members, and the many participants who provided lively discussions and shared additional information and references during the workshop. We would especially like to thank Leslie Saldias for organising the logistics of the event.

References

Allen, A. 2021, arXiv:2111.12574
Arviset, C. et al. 2021, LPI Contributions, 2549, 7014
Bonaldi, A. et al. 2021, MNRAS, 500, 3821
Major, B. et al. 2019, ASP Conf. Ser., 523, 277
Nordin, J. et al. 2019, A&A, 631, A147

Links

[1] Workshop programme: https://www.eso.org/sci/meetings/2022/REPRODUCIBILITY2022/programme.html
[2] ESA datalabs: https://datalabs.esa.int/
[3] Spanish Virtual Observatory: https://svo.cab.inta-csic.es/main/index.php
[4] CANFAR: https://www.canfar.net/en/
[5] SKAO: https://www.skao.int/en
[6] DORA: https://sfdora.org/
[7] ASCL: https://ascl.net/
[8] SRON reproduction package template: https://github.com/jdeplaa/open-data-template
[9] Docker container: https://www.docker.com/
[10] Singularity container: https://docs.sylabs.io/guides/3.5/user-guide/index.html
[11] Maneage: https://maneage.org/
[12] ESCAPE: https://projectescape.eu/
[13] ESO Phase 3 archive: https://www.eso.org/sci/observing/phase3.html
[14] ODYSSEUS: https://sites.bu.edu/odysseus/
[15] Conference YouTube page: https://www.youtube.com/channel/UC34OkFIBFtCHpjkNOOM7HOA/videos



ESO/Gerhard Hüdepohl

ESO’s Very Large Telescope (VLT) seen from above.