1st Conference on Research Data Infrastructure

Enabling RDM

https://doi.org/10.52825/CoRDI.v1i.331

Published: 07 Sept. 2023

Born-fair data projects using cookiecutter templates

Felix Henninger[1][https://orcid.org/0000-0002-7730-9511]

[1] Ludwig-Maximilians-Universität, Germany

Keywords: RDM, Data stewardship, FAIR data principles, Research Software, BERD@NFDI

Implementing research data management best practices and fair principles (Wilkinson et al., 2016) is vital for transparent, reproducible research, as well as efficient, sustainable science that avoids duplication of effort. Scientists can also benefit directly from incorporating data management into their data collection and analysis workflows. However, there is an initial cost to adoption that poses a burden and substantial barrier to entry even to well-intentioned researchers. In our experience in statistical and rdm-focused consulting, this cost increases as a project progresses, with a late-stage conversion being the most costly in terms of resources and energy because a working analysis needs to be adapted in one step. Therefore, we believe that it is useful to adopt best practices for data stewardship early-on in a project, and ideally from the get-go. In this contribution, we present a tool for creating and instantiating project templates that conform to good practices with regard to the data management and analysis projects more generally. In the same vein as “born-open data” (Rouder, 2016) where data is published immediately upon collection, our goal is to establish born-fair datasets that implement proven methods for data stewardship as early on in the research datalifecycle as is feasible. Our aim is to encourage researchers and analysts to incorporate best practices into their workflows from the onset by providing data and analysis templates that implement desirable properties. By adopting these templates, researchersimmediately gain access to a number of tools that simplify their work and make it more efficient, while also providing a foundation for increased reproducibility, data documentation through codebooks, as well as metadata for long-term archival. Because abroad-strokes approach may not work in practice due to idiosyncrasies of individual research projects, one size may not fit all. For this reason, the templates contain customisation options that researchers can use to tailor the templates to their requirements.

The templates build on the well-established cookiecutter library for the Python programming language (Greenfeld et al., 2022), which we additionally extend to R, aprogramming language somewhat more common among statisticians and social scientists, thereby creating a cross-platform infrastructure. Both libraries create a project skeleton with a pre-specified directory structure, and include configuration for commonly used tools. Upon template creation, a wizard guides users through a customisation step, allowing them to adapt templates to their needs and catering to the demands of a project at hand.

Owing to the open-source nature of the project and the firmly established and well-documented standard, researchers can easily adapt templates and create their own, to accommodate their specific needs and domain requirements. We hope to foster a community of researchers who share and improve their workflows, and anticipate further uses of the templates for teaching and other purposes.

At CoRDI, we hope to introduce our project to the wider nfdi community and propose it as a lightweight, interoperable and interdisciplinary standard, benefiting all researchers across domains. By streamlining advanced users' workflows, and making reproducible practices more accessible, we aim to enable and facilitate the uptake of rdm across the communities represented there, and build integrations to interoperate with the multitude of services currently under development and in use.

To summarise, we introduce templates for data analysis and archival that researchers can apply themselves, to render possible and encourage better practices during analysis, and prepare data for long-term storage and later re-use. Our hope is to encourage more researchers to adopt rdm best practices more frequently and earlier in projects, demonstrating the value of a more structured workflow and facilitating a shift to FAIR principles mor generally

References

1. Greenfeld, A. R., Greenfeld, D. R., Pierzina, R., & Contributors (2022). Cookiecutter (Version 2.1.1) [Computer software]. GitHub. https://github.com/cookiecutter/cookiecutter/
2. Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data, 3(1), Article 1. https://doi.org/10.1038/sdata.2016.18