A FAIR and project-oriented template for open science data workflows

Kauê de Sousa[a,b,*], Marie-Angélique Laporte[c]

[a] *Digital Inclusion, Bioversity International, Montpellier, France*
[b] *Department of Agricultural Sciences, University of Inland Norway, Hamar, Norway*
[c] *Performance Innovation and Strategic Analysis for Impact (PISA4), Bioversity International, Montpellier, France*

Abstract

Managing research data in accordance with the FAIR principles (Findable, Accessible, Interoperable, and Reusable) is essential for ensuring transparency, reproducibility, and effective scientific collaboration (Wilkinson et al., 2016). Reproducibility, the ability to repeat an analysis, and replicability, the ability to repeat an experiment under similar conditions (Stevens, 2017), are foundational to trustworthy and collaborative science (Powers and Hampton, 2019; Munafò et al., 2017). These practices enable researchers to revisit analyses after extended periods and allow peers to verify results or generate new insights from existing data. In this context, we present a structured and reusable repository template specifically designed to support FAIR data workflows and promote both reproducibility and replicability across research projects.

Keywords: data management, metadata, FAIR principles, reproducibility, open science

Suggested citation

de Sousa, K.; Laporte, M.-A. (2025) A FAIR and project-oriented template for open science data workflows. `https://hdl.handle.net/10568/174310`.

*Corresponding author
Email addresses: `k.desousa@cgiar.org` (Kauê de Sousa), `m.a.laporte@cgiar.org` (Marie-Angélique Laporte)

Statement of need

Data-driven research faces several challenges, including inadequate data documentation, inconsistent structure, and poor reproducibility (Wilkinson et al., 2016; Munafò et al., 2017). Researchers often struggle with these aspects, lacking practical examples or accessible tools. To address these issues, we created a project-oriented, easy-to-use data template. It streamlines metadata management, facilitates reproducibility, and ensures alignment with FAIR principles.

Repository structure

The template repository includes clear, intuitive folder structures and essential resources.

```
template-repo-data-analysis/
|-- data/                                       # Anonimized raw and cleaned datasets
|-- docs/                                       # Reports or additional documentation
|-- metadata/                                   # Metadata files and templates
|    |-- project-metadata.xlsx                  # Project metadata
|    |-- project-metadata.json                  # Project metadata for DataCite
|    |-- example-metadata-data-mip-uganda.csv   # Description of the dataset (example)
|    |-- README.md
|-- output/                                     # Model results, figures, tables
|-- script/                                     # Data analysis scripts
|-- .gitignore                                  # List of files to exclude from Git
|-- LICENSE                                     # License stablishing the rights to use
|-- CITATION.cff                                # Used by Zenodo to generate citation
|-- template-repo-data-analysis.Rproj           # RStudio file to set up the environment
|-- README.md                                   # Project overview
```

Key features

- FAIR-aligned metadata: Comprehensive metadata files (.xlsx, .json) aligned with Zenodo (`https://zenodo.org`) and DataCite (Robinson-Garcia et al., 2017) schemas.
- Automated workflows: R scripts (R Core Team, 2024) to validate and convert metadata, ensuring consistent quality.
- Citation integration: A CITATION.cff file for automatic citation generation.
- Reproducibility: Scripts and clear instructions ensure any researcher can reproduce results.
- Open and collaborative: Open licensing (CC BY 4.0, `https://creativecommons.org`) promotes collaboration, reuse, and attribution.

Implementation and usage

Researchers can easily adopt this repository by:

1. Cloning the repository from GitHub (`https://github.com/AgrDataSci/template-repo-data-analysis`).
2. Adding their data and performing their analysis.
3. Filling in the provided project-metadata.xlsx template.
4. Running provided R scripts for validation and metadata generation.
5. Publishing via integrated platforms like Zenodo for long-term archiving and DOI assignment.

Conclusion

Adopting structured, FAIR-aligned workflows can significantly enhance reproducibility, collaboration, and transparency in research. This template provides a practical starting point, promoting best practices from the outset of research projects.

Acknowledgments

This work has been supported by the projects 1000FARMS (INV-031561) funded by the Gates Foundation, and BOLD Work Package 7 (Crop Trust Ref: CONT-1522) funded by the Norwegian Agency for Development Cooperation (Norad) through The Global Crop Diversity Trust.

References

Marcus R Munafò, Brian A Nosek, Dorothy V M Bishop, Katherine S Button, Christopher D Chambers, Nathalie Percie du Sert, Uri Simonsohn, Eric-Jan Wagenmakers, Jennifer J Ware, and John P A Ioannidis. A manifesto for reproducible science. *Nature Human Behaviour*, 1(1):21, 2017. ISSN 2397-3374. doi: 10.1038/s41562-016-0021. URL `https://doi.org/10.1038/s41562-016-0021`.

Stephen M Powers and Stephanie E Hampton. Open science, reproducibility, and transparency in ecology. *Ecological Applications*, 29(1):e01822, jan 2019. ISSN 1051-0761. doi: 10.1002/eap.1822. URL `https://doi.org/10.1002/eap.1822`.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2024. URL `https://www.R-project.org/`.

Nicolas Robinson-Garcia, Philippe Mongeon, Wei Jeng, and Rodrigo Costas. Datacite as a novel bibliometric source: Coverage, strengths and limitations. *Journal of Informetrics*, 11(3):841–854, August 2017. ISSN 1751-1577. doi: 10.1016/j.joi.2017.07.003. URL `http://dx.doi.org/10.1016/j.joi.2017.07.003`.

Jeffrey R. Stevens. Replicability and Reproducibility in Comparative Psychology. *Frontiers in Psychology*, 8:862, may 2017. ISSN 1664-1078. doi: 10.3389/fpsyg.2017.00862. URL `http://journal.frontiersin.org/article/10.3389/fpsyg.2017.00862/full`.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. The fair guiding principles for scientific data management and stewardship. *Scientific Data*, 3(1), March 2016. ISSN 2052-4463. doi: 10.1038/sdata.2016.18. URL `http://dx.doi.org/10.1038/sdata.2016.18`.