

Open science and reproducibility

The FINEPRINT infrastructure

Authors: Victor Maus, Nikolas Kuschnig & Stefan Giljum

Institute for Ecological Economics, Vienna University of Economics and Business (WU)

FINEPRINT Brief No. 1, November 2018

With the ongoing reproducibility crisis casting doubts on scientific findings, we follow strict transparency and reproducibility in the FINEPRINT project, applying state-of-the-art, open tools and best practice research guidelines.

Introduction

Reproducing an experiment is essential for scientific development. Recent publications have shown a high level of concern about irreproducibility of results [1–4]. A survey published in *Nature* [1], involving more than 1,500 researchers, showed that more than 70% of them have failed to reproduce other scientists' experiments and more than half have failed to replicate their own experiments. This has led to discussions on how researchers, journals, institutions, and funding agencies can increase transparency and reproducibility.

In the FINEPRINT project we implement good practice guidelines and tools to ensure the reproducibility of our analyses and publications. We want to share our work with the world openly, as we believe that transparency is the best way to build trust in scientific findings. Furthermore, we hope to speed up scientific discoveries by allowing everybody to contribute, improve methods and answer new research questions.

Ensuring the reproducibility of our research findings means keeping track of a variety of data, scripts, models and software versions. This includes handling several heterogeneous datasets derived from spatial and non-spatial data sources (e.g. global crop maps and commodity trade data) as well as versions of our scripts and models (e.g. [FABIO](#)). Furthermore, open distribution and proper documentation of research procedures are essential to achieve reproducibility. To improve our workflow, we have selected a set of open source tools and adopted best practice guidelines. In this Brief, we introduce our infrastructure, implemented to support our goals towards reproducibility and open science.



Infrastructure

A large part of our research is done before performing actual data analysis or modelling, for example, when investigating data availability or different methodological approaches. This part of the scientific process is vital and worth proper documentation to facilitate the reproduction of findings. We adopted [OpenProject](#) as a web-based tool for task and project management and built a web service for research documentation using R Markdown. These tools facilitate collaboration and transparency across the team and always provide the most recent version of work and a complete version history. This setup ensures that our team always stays up-to-date regarding progress and future working steps (see Figure 1).

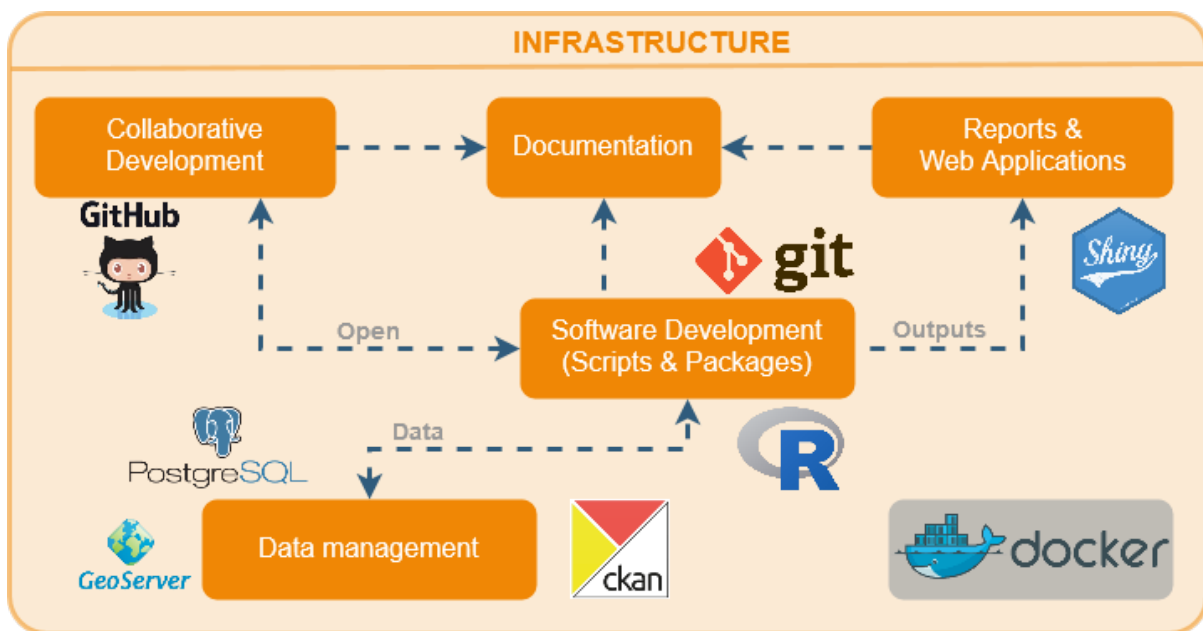


Figure 1: Outline of the FINEPRINT infrastructure

We adopted [Git](#) and [GitHub](#) to facilitate collaborative code development, adequate version control as well as a platform to share our work with the scientific community. The FINEPRINT repositories are available online at github.com/fineprint-global. As we implement most of our work in R, we introduced best practice coding guidelines, along with the style guide proposed by [5]. Besides enforcing reproducibility, these guidelines help to improve the readability of the code and scripts - both for other researchers and ourselves

The execution of the FINEPRINT project requires handling vast amounts of data, especially in geospatial formats. Therefore, we implemented data management systems that can properly handle data and metadata and provide easy access for processing and visualization. Our system combines [Geoserver](#), for geospatial data management, and [CKAN](#), an open data website tool. This infrastructure allows us to store, search, preview and access vast amounts of spatial and non-spatial data via web interface, web data services (e.g. using [OGC Standards](#) and through an application programming interface (API) from within in our scripts.

Keeping track of our own development is not sufficient for ensuring reproducibility, due to frequent updates to software dependencies. We use [Docker](#) to provide a reproducible computational



environment and facilitate the management of applications. Docker is a software platform to build and manage applications in an isolated computational environment - namely in containers, which carry all software dependencies and versions; i.e. akin to virtual machines but usually lighter and faster. The Docker engine provides the foundation for the dynamic and portable setup of several applications currently running on our server.

Finally, we are developing interactive data visualizations using [Shiny](#), in order to allow website users to customise graphical analyses and download corresponding data. We have already developed similar visualisation tools on other platforms - for example [materialflows.net](#) for visualising global material flow data.

Conclusion

The setup of the FINEPRINT infrastructure provides us with a capable and flexible computational environment to support the reproducibility and transparency of our research. Datasets can be added to our services manually or via script, previewed, searched and retrieved through an intuitive web interface or via API. Extensible services and a flexible setup allow us to prepare for and adapt to future challenges, for example scaling processing power to analyse large datasets.

By making our research reproducible and open we aim to build trust in scientific findings. We also expect speeding up scientific discoveries by facilitating the engagement of other research teams in improving methods and answering new research questions in the field of sustainability science.

Citation

Maus, V., Kuschnig, N., Giljum, S. 2018. Open science and reproducibility. The FINEPRINT infrastructure. FINEPRINT Brief No. 1. Vienna University of Economics and Business (WU). Vienna.

References

- [1] Baker M. 1,500 scientists lift the lid on reproducibility. *Nature* 2016;533:452-4. doi:[10.1038/533452a](#).
- [2] Buck S. Solving reproducibility. *Science* 2015;348:1403-3. doi:[10.1126/science.aac8041](#).
- [3] Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? *Science Translational Medicine* 2016;8:341ps12-2. doi:[10.1126/scitranslmed.aaf5027](#).
- [4] Camerer CF, Dreber A, Holzmeister F, Ho TH, Huber J, Johannesson M, et al. Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour* 2018;1. doi:[10.1038/s41562-018-0399-z](#).
- [5] Wickham H. *R packages: Organize, test, document, and share your code*. O'Reilly Media, Inc. 2015.