Caring is Sharing – Exploiting the Value in Data for Health and Innovation
M. Hägglund et al. (Eds.)

doi:10.3233/SHTI230064

Reproducibility in 2023 - An End-to-End Template for Analysis and Manuscript Writing

Jonathan M. MANG[a,1], Hans-Ulrich PROKOSCH[a,b] and Lorenz A. KAPSNER[a,c]
[a] *Medical Center for Information and Communication Technology, Universitätsklinikum Erlangen, Erlangen, Germany*
[b] *Chair of Medical Informatics, Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU), Erlangen, Germany*
[c] *Institute of Radiology, Universitätsklinikum Erlangen, Friedrich-Alexander-University Erlangen-Nürnberg (FAU), Erlangen, Germany*

Abstract. Reproducibility imposes some special requirements at different stages of each project, including reproducible workflows for the analysis including to follow best practices regarding code style and to make the creation of the manuscript reproducible as well. Available tools therefore include version control systems such as Git and document creation tools such as Quarto or R Markdown. However, a re-usable project template mapping the entire process from performing the data analysis to finally writing the manuscript in a reproducible manner is yet lacking. This work aims to fill this gap by presenting an open source template for conducting reproducible research projects utilizing a containerized framework for both developing and conducting the analysis and summarizing the results in a manuscript. This template can be used instantly without any customization.

Keywords. Reproducibility, Data sharing, Template, Container, Computer and information sciences

1. Introduction

The goal of Data Science is to generate new insights from data and encompasses a wide spectrum of data handling procedures, from the initial data collection further on to data processing and preparation, as well as data exploration and statistical modeling, finally to visualization, interpretation and publishing the results [1]. The breadth and complexity of data science scenarios requires comprehensive expertise and experience in a wide range of related topics and domains, e.g. mathematics, statistics, computer science, software development, and domain knowledge of the problem to solve, besides others. Simultaneously, there is increasing interest in automating parts of the data science process, if not the entire process [1]. Additionally, researchers from a number of disciplines in the computer sciences call for *reproducibility* or *reproducible research* as a minimum achievable standard for assessing the value of scientific claims, especially when full independent replication of a study is not possible, for example, due to lack of resources such as time and money [2–5].

[1] Corresponding Author: Jonathan M. Mang, E-mail: jonathan.mang@uk-erlangen.de

Reproducibility in this context refers to obtaining consistent results using the same input data, calculation steps, methods, and codes, as well as the same analysis conditions [6]. As early as 1992, there were detailed ideas and concepts for realizing the goal of reproducible science in which authors are able to link the underlying data, parameters, and programs to each figure in the manuscript so that recalculations can be initiated with the click of a button [7]. In addition, the constantly increasing availability of (big) data coupled with rising computing power is enabling ever more complex, automated as well as multi-site analyses [8–10]. Although the intention to share data and code openly is widely established and a variety of methods and tools have been published to accomplish reproducibility in all steps of scientific research, an easy-to-use and publicly available project template is yet missing to organize reproducible end-to-end analysis including the final manuscript for publishing the results.

Since reproducible deployments are common in software development and the creation of reproducible reports is common in the statistical area, these concepts should be merged and the runtime environment in which the analysis was performed should also be provided in a reproducible manner to ensure the reproducibility of the entire workflow besides the publication of data and code.

This project template provides such a framework in which both the execution of the analysis and creation of the manuscript remain reproducible, including the entire software environment required and its dependencies using state-of-the-art tools.

2. Methods

Many tutorials, tools, and reports are available to facilitate approaching the topic of reproducibility in science [5,11,12]. These reports present in detail tool collections and concepts that can be used within the different stages of an analysis: A good data management is useful to structure all documents and materials as a basis for effective planning of the project which is addressed with a predefined folder structure in the template. Since not only the documents and files, but also the program code should be neat and traceable, as well as regularly backed up to different locations, GitHub was used for decentralised versioning the program code, files and folders in the presented template.

Intermediate steps of an analysis should be clearly traceable with intermediate results [11]. Therefore, all raw data is available in a usable form (accessible, digital and non-proprietary) within the template. Ideally, the rationale for decisions made during the analysis should be recorded and traceable as well. To this end, a small, annotated demo analysis has been integrated into the template, describing the data flow in the code up to the final calculation results embedded in the manuscript. The code follows a clean, consistent style that is intended to make the code easier to read. Additional repetitive tasks are automated using functions.

Some published methods and guidance on how reproducible research is possible also consider the benefits of containerizing the analysis [11]. Thus, with appropriate encapsulation of the scripts and the development environment, it can be ensured that not only the text, the underlying data and the scripts necessary to reproduce the results are available, but also that the complete computational and system environment, including all system and package dependencies, can be used in a defined, versionable condition to

reproduce the results independently of the local system configuration and time. In the provided template, Docker [13,14] was used to implement containerization.

Even after analysis, the container-based setup can be used to summarize the results in a reproducible manuscript format, eliminating the need for manual copying and pasting of data into text or tables, or manual creation of figures. From data analysis to a dynamic document embedding the results, there are many tools to support the publication workflow. In the following, we use Quarto [15], a modern open source scientific and technical publishing system built on Pandoc. Various languages such as R [16] and Python [17] can be used in conjunction with Markdown [18], to create a publication-ready manuscript while adhering to accepted standards.

3. Results

At first, it is sufficient to have Docker [13] and Git [19] installed locally to use the presented framework. The repository of the template (see section "Declarations") contains the structure for an ordered data storage, the code necessary for the analysis, as well as a prepared and directly usable container image, in which the analysis can be performed directly, independent of the local operating system. The container image provided ensures that all packages and system requirements necessary to perform the analysis and display the results are uniformly bundled and identical for each execution.

Next, when analyzing the data, the containerized development environment *RStudio* [20] provides an interface to various programming languages as well as standard code development tools such as auto-completion, syntax highlighting, a variable browser and plug-ins, as well as publishing tools (see Figure 1).

The analysis code, the manuscript (including the literature file) and the container information are part of the code repository and are therefore version controlled. Collaboration with co-authors is easily managed using common Git-based code management tools such as branches and merge requests.

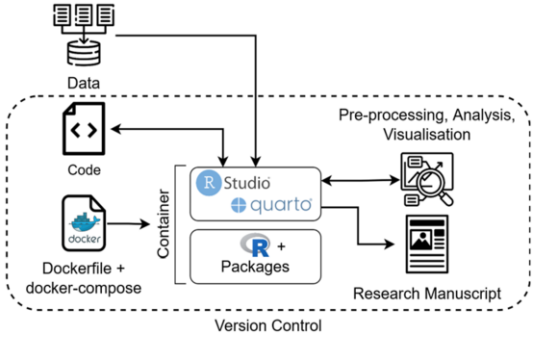

Figure 1. Exemplary workflow to ensure reproducibility in data analysis and manuscript writing. The code for pre-processing, analysis and presentation of results in the manuscript ('code') is version controlled. The build information of the container for the runtime and development environment ('Dockerfile + docker-compose') is also version controlled. The container image is also version controlled and stored in a dedicated online repository. The raw data itself is not version controlled via Git.

Finally, once the analysis is complete, details such as the correct formatting of sources and rendering into the correct output format of the target journal need to be done. For this purpose, Quarto offers a way to integrate journal templates [21] or journal-

specific citation styles [22] without having to change the manuscript text. Once all the details used for the analysis and evaluation as well as the Dockerfile required to create the container with the underlying runtime environment have been stored in the repository, they are shared publicly via the version control system used. To enable reproducibility, the container image with all dependencies is stored in a publicly available package repository. Further details and instructions on all the necessary steps are described in the template.

4. Discussion

We here present a project template that includes currently available tools that are required for conducting reproducible research. Besides the version control of the analysis code and the writing of the manuscript, reproducibility is here extended also to the runtime environment in which the former are performed. The presented template uses GitHub [23] for version control, R [16] as the main programming language, RStudio [20] as the corresponding development environment, and Docker [13] for containerization. Nevertheless, further technologies may be used instead.

Completely reproducible analyses requires all data to be available so that the same results can then be obtained with the analysis steps, which need to be published as well. However, often the raw data cannot be published due to various reasons (storage requirements, data protection, trade secrets). In this case, aggregated data or synthetic data with a structure identical to the original data could be provided instead.

One way to protect intellectual property may be to place restrictions on the permissible use of the dataset in a license or as part of a formal agreement between the researcher and an applicant [3].

5. Conclusions

The purpose of this article is to summarize current state-of-the-art technologies to ensure reproducibility in research and to provide a ready-to-use project template for this purpose. In addition to the methods described for publishing the data and code, this template provides a framework for extending the concepts of reproducibility to the summary of the final results in the form of a manuscript. For this purpose, both a Quarto template that connects the manuscript text with the analysis code and dynamically embeds the results as well as the underlying runtime environment in the form of a Docker container are provided. Together with the publicly available open source software framework used for this purpose, this enables timeless reproducibility of research results.

Declarations

Availability of data and materials: The template repository repub is available on GitHub (https://github.com/joundso/repub). Extensive information about the usage and optional parameterization of the template is available in the respective readme files in the repository.

Funding: This work was partly funded by the German Federal Ministry of Education and Research (BMBF) within the Medical Informatics Initiative (MIRACUM Consortium) under the Funding Number FKZ: 01ZZ1801A.

Authors' contributions: JMM and LAK set up the template. JMM wrote the original manuscript. LAK and HUP supervised the project, reviewed, and edited the manuscript. All authors read and approved the final version. The present work was performed in (partial) fulfillment of the requirements for obtaining the degree "Dr. rer. biol. hum." at Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (JMM).

Competing interests: None declared.

References

[1] De Bie T, De Raedt L, Hernández-Orallo J, Hoos HH, Smyth P, Williams CK. Automating data science. Communications of the ACM. 2022 Feb 23;65(3):76-87.
[2] Schwab M, Karrenbach N, Claerbout J. Making scientific computations reproducible. Computing in Science & Engineering. 2000 Nov;2(6):61-7.
[3] Laine C, Goodman SN, Griswold ME, Sox HC. Reproducible research: moving toward research the public can really trust. Annals of Internal Medicine. 2007 Mar 20;146(6):450-3.
[4] Gentleman R. Reproducible Research: A Bioinformatics Case Study, Statistical Applications in Genetics and Molecular Biology. 4 (2005). doi:10.2202/1544-6115.1034.
[5] Peng RD. Reproducible Research in Computational Science, Science. 334 (2011) 1226–1227. doi:10.1126/science.1213847.
[6] National Academies of Sciences, and Medicine, Reproducibility and Replicability in Science, The National Academies Press, Washington, DC, 2019. doi:10.17226/25303.
[7] Claerbout JF, Karrenbach M. Electronic documents give reproducible research a new meaning. InSEG technical program expanded abstracts 1992 1992 Jan 10 (pp. 601-604). Society of Exploration Geophysicists.
[8] Prokosch HU, Bahls T, Bialke M, Eils J, Fegeler C, Gruendner J, Haarbrandt B, Hampf C, Hoffmann W, Hund H, Kampf M. The COVID-19 data exchange platform of the German university medicine. InChallenges of Trustable AI and Added-Value on Health 2022 (pp. 674-678). IOS Press.
[9] Kapsner LA, Kampf MO, Seuchter SA, Gruendner J, Gulden C, Mate S, Mang JM, Schüttler C, Deppenwiese N, Krause L, Zöller D. Reduced rate of inpatient hospital admissions in 18 German university hospitals during the COVID-19 lockdown. Frontiers in public health. 2021:1018.
[10] Schüttler J, Mang JM, Kapsner LA, Seuchter SA, Binder H, Zöller D, Kohlbacher O, Boeker M, Zacharowski K, Rohde G, Balig J. Letalität von Patienten mit COVID-19: Untersuchungen zu Ursachen und Dynamik an deutschen Universitätsklinika. Anästhesiologie & intensivmedizin. 2021;62:244-57.
[11] Alston JM, Rick JA. A beginner's guide to conducting reproducible research. Bulletin of the Ecological Society of America. 2021 Apr 1;102(2):1-4.
[12] Sandve GK, Nekrutenko A, Taylor J, Hovig E. Ten simple rules for reproducible computational research. PLoS computational biology. 2013 Oct 24;9(10):e1003285.
[13] Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux j. 2014 Mar 2;239(2):2.
[14] Bernstein D. Containers and cloud: From lxc to docker to kubernetes. IEEE cloud computing. 2014 Sep;1(3):81-4.doi:10.1109/MCC.2014.51.
[15] Quarto, RStudio, PBC, 2022. https://quarto.org/ (accessed December 5, 2022).
[16] R Core Team, R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, 2022. https://www.R-project.org/.
[17] G. Van Rossum, and F.L. Drake Jr, Python reference manual, Centrum voor Wiskunde en Informatica Amsterdam, 1995.
[18] Daring Fireball: Markdown Syntax Documentation, (n.d.). https://daringfireball.net/projects/markdown/syntax#philosophy (accessed December 27, 2022).
[19] S. Chacon, and B. Straub, Pro git, Apress, 2014.
[20] Posit team, RStudio: Integrated development environment for R, Posit Software, PBC, Boston, MA, 2022. http://www.posit.co/.
[21] Quarto Journal Templates, (2023). https://github.com/quarto-journals (accessed January 9, 2023).
[22] Citation Style Language, (2023). https://citationstyles.org/ (accessed January 9, 2023).
[23] github, GitHub, (2020). https://github.com/.