Website: https://osf.io/phmce/
Version: July 2019
DOI: 10.17605/OSF.IO/CB7Z8

AGILE Association of Geographic Information Laboratories in Europe

REPRODUCIBLE PAPER GUIDELINES

Full and short paper submissions to the AGILE conference **must** include a **Data and Software Availability** sub-section as part of the Methods section. The section documents all data, software, and computational infrastructure to support reproduction, or otherwise mentions reasons for not publishing them.

PRE-SUBMISSION REPRODUCIBILITY CHECKLIST

For all **datasets** included/produced in the submission, check if:

- ❑ Data is provided in a non-proprietary format (if necessary, export from proprietary format for publication)
- ❑ Data is documented (at least description of collection query and field or column names, ideally using complete metadata following established standards)
- ❑ Data is accessible in a public repository
- ❑ Data has a clear licence

For any **software tool/library/package** used or produced, check if:

- ❑ Computational environment (including hardware) is documented or provided in the most appropriate format given its complexity
- ❑ The versions of relevant software components (libraries, packages) are provided
- ❑ Software is available in a public repository
- ❑ Software has a clear license
- ❑ Computational steps are explained in a text file, flowchart, or script
- ❑ All parameters needed to run the computational workflow are provided

In the **Data and Software Availability section**, check if you include:

- ❑ Data and software statements according to the template
- ❑ The reasons, if any, for not being able to share (parts of) data or code.

For properly **acknowledging data and software** by both you and others check that:

- ❑ All datasets and code used or mentioned are cited throughout the paper and included in the references with DOIs.

WRITING THE DATA AND SOFTWARE AVAILABILITY SECTION

The Data and Software Availability section provides references to where data and software can be accessed, where the software documentation is available (e.g. other article section, or README file in repository) and under what conditions. It should be concise and contain links to archival repositories as persistent links, ideally DOIs. Possible conditions include copyright, licenses, or access procedures for protected data. In addition to the direct links within this section, you should cite data and software as needed elsewhere in the paper. Ensure that all the datasets and (third-party) software tools used in the paper are cited as requested by the respective projects. If a project does not provide a recommended citation or reference then follow standard citation guidelines.

Possible statements which you might include in the Data and Software Availability sub-section are provided below. You may blank links for a double-blind peer review process, or use anonymous access links if your repository supports them.

All research data supporting this publication ...

… are available in [add the name of the repository(-ies)] under the following DOI [add a DOI link(s);]

… was accessed on [add date of dataset access/download] *with the following query parameters* [add query parameters, if applicable] *under the license* [indicate dataset license].

The computational workflow in this publication …

… is executed via a single script file published under license [the license] *at* [DOI of repository].

… is published in a [language] *module/package at* [link of software project]*, the used version in archived at* [DOI of repository].

… is provided as a [container/VM] *published at* [DOI of repository]*, with instructions included in the file README.txt in the repository.*

Research data/code is not publicly available due to [indicate reasons, e.g. licenses, privacy statements; if there are processes to obtain the data, describe them].

RATIONALE

Authors should publish all parts of their computational workflow in a place ensuring long-term accessibility and in a format enabling reproduction and re-use. Every step toward reproducibility is better than not publishing anything. While open access to data and code is an asset for enabling reproducibility, reasons for refraining from making (parts of) the material openly available need to be disclosed. The structure (file organisation and naming conventions) and documentation (e.g. readme files) of data, software, and computational infrastructure should be sufficiently clear to allow readers to recreate the original authors' computational environment as closely as possible.

Our guidelines focus strictly on computational reproducibility: that the data and software be provided in a way that the analysis may be re-run to reproduce the study results[1]. While it is also critical that experimental protocols and hardware (for example, the model numbers and specifications of sensors used to gather data) be described in sufficient detail that an independent scientist may replicate the study, such considerations are outside the scope of these guidelines.

The following guide targets authors and reviewers. The recommendations span from the minimum requirements to ideal aspirations, and are illustrated by concrete examples from the geoinformatics domain. In some cases, we give examples for intermediate steps towards higher reproducibility.

MOTIVATION

Research in geoinformatics and GIScience frequently involves data and computational methods. Despite the expected high standards of documentation of work in a paper, textual contributions fall short in allowing readers to assess the computational aspects of research work. AGILE's Reproducible Paper Guidelines are part of AGILE's mission to promote research, education and networking of the geoinformatics community in Europe and beyond.

The objective of providing data and software for reproduction, through third parties or the original authors, improves the documentation of computational research and enables better science.

It also creates opportunities for increased interaction within the community: methods developed can be tested and applied to further study areas, research can be used in an educational context, and in-depth feedback on research work can be provided.

Authors receive recognition not only in regards to the body of literature, but also in regards of a body of data, code and best practices.

VISION

The primary maxim for everyone involved in creating, reviewing and interacting with manuscripts is: aim for the maximum level of reproducibility and be supportive and kind in all interactions. All principles and recommendations in these guidelines are to be used to promote reproducibility and foster collaboration, and never to exclude or discriminate.

The AGILE community embraces reproducibility and provides a positive setting for researchers of all career stages, skill levels, and backgrounds to face the challenges of reproducibility. In the long run, all conference contributions become reproducible, and AGILE research lab members effectively achieve reproducibility by internalising its advantages and principles and reap its rewards in quality research.

[1] Definition of reproducibility followed in these guidelines in the Claerbout/Donoho/Peng terminology, cf. Barba, L.A. (2018). Terminologies for Reproducible Research. CoRR, abs/1802.03311 (arXiv:1802.03311).

AUTHOR GUIDELINES

DATA IN RESEARCH PAPERS

	Minimum	Ideal
What?	Publish all input data + data description / documentation	Publish all data and adhere to standardised, discipline-specific metadata[2] to describe your data
Where?	Use a data repository providing a DOI[3]	Use a discipline-specific repository[4] with a DOI
How?	Use open data formats + specify a license	Make your data FAIR (Findable, Accessible, Interoperable and Reusable) and as open as possible

What if...

- the datasets are **already openly available**, especially with a DOI (digital object identifier), then cite the dataset[5], and clearly indicate which subset (if any) has been used.
- the datasets you used **are not openly available**, or only temporarily available or it is too difficult to recreate the subset used then - if the original dataset license permits - **upload the dataset into a repository with a DOI.**
- **the licence, or privacy considerations, do not permit public re-sharing** of the dataset, or parts of it, then document the dataset and explain the procedures and conditions needed to access it. Provide a synthetic dataset to demonstrate your workflow.
- **you are the creator** of the dataset, then give it a license that allows the broadest possible reuse.
- your data is published under your name already, then you can use **anonymised links** in repositories like OSF[6] or FigShare[7] to support blind review.

Examples

- **Social media data**: If the platform's terms of service do not allow for sharing all the data in a repository, as a minimum, provide unique identifiers of the posts used[8].
- **OpenStreetMap data**: as a minimum, provide feature type(s) used, geographic coverage, and the date of extraction or usage, ideally upload the extract to a data repository.
- **Framework data, socio-demographic and statistical data** (e.g administrative or natural boundaries, elevation data, 3D city models): use the appropriate unique identifier to cite the dataset, e.g URI, DOI, POI, and describe the exact data source and the timestamp.
- **Personal data** (data containing information which can lead to the identification of individuals) should be shared after anonymisation / sufficient aggregation. If this is not possible, dataset can be uploaded to a restricted access repository (e.g. DANS Easy) and metadata can be made public.
- **Scraped data from websites** (e.g. real estate values, sports tracking applications): if the platform's terms of service do not allow for sharing all the data in a repository, as a minimum provide metadata and scraping script with all its parameters.

Find more and updated examples in the guidelines wiki: https://osf.io/phmce/wiki/

[2] Metadata standards catalogue: https://rdamsc.bath.ac.uk/
[3] https://en.wikipedia.org/wiki/Digital_object_identifier
[4] Registry of research data repository: https://www.re3data.org/, for example for earth sciences: https://www.pangaea.de or for datasets in netCDF format: https://researchdata.4tu.nl/en/home/
[5] Data Citation Principles: https://www.force11.org/datacitationprinciples
[6] Anonymous dataset for peer review in OSF: http://help.osf.io/m/links_forks/l/524049-create-a-view-only-link-for-a-project
[7] Guidance how to create anonymous dataset for peer review in Figshare and Zenodo: https://ineed.coffee/5205/how-to-disclose-data-for-double-blind-review-and-make-it-archived-open-data-upon-acceptance/
[8] Report on preserving social media data: https://www.dpconline.org/docs/technology-watch-reports/1486-twr16-01/file

COMPUTATIONAL WORKFLOWS IN RESEARCH PAPERS

	Minimum	Intermediate	Ideal
What? Computational environment	Describe the environment and computational infrastructure, e.g. computer specs, operating system + software versions	Provide live documents (structured configuration files with dependency information, e.g. a Binder[9])	Provide the actual environment, e.g. a container created by a Dockerfile[10] or a Virtual Machine (VM, e.g. OSGeo-Live)
Computation steps	Document the detailed steps in a text file and/or flowchart (every action/click)	Provide scripts / models and a README file that explains their use	Provide a software package with structured metadata[11], tests/CI[12], and an automated workflow[13] + If applicable: Add link to running instance of software
Where?	Repository providing a DOI, such as Zenodo, OSF, b2share, or FigShare		Minimum + versioned code repository, such as GitHub or GitLab
How? Tools used	Use generally available proprietary tools (avoid tools that are not available to reviewers and other researchers)		Use (and create) open source tools; cite core modules/tools/language used, including your own
Development practices	Use clear licenses[14] that fit your environment	Follow "Good enough practices" for scientific computing software[15]	Use development guidelines for your environment / language of choice (e.g. for R[16])

Examples

- **Live documents:** run code from GitHub in online notebook e.g. Binder
  - Python: https://github.com/jorisvandenbossche/geopandas-tutorial
  - R: https://github.com/nuest/sensebox-binder (https://zenodo.org/record/1139929)
- **Repository structure**
  - R: https://github.com/Reproducible-Science-Curriculum/rr-init and https://github.com/benmarwick/rrtools
- **Container with computational environment**
  - Docker-based Geospatial toolkit for R: https://github.com/rocker-org/geospatial
  - OSGeo-related Docker images: https://wiki.osgeo.org/wiki/DockerImages
  - Examples of Research Compendia / reproducibility packages on Zenodo: https://zenodo.org/communities/research-compendium?page=1&size=20
- **AGILE example paper:** Keßler & Lotstein (2018)[17]

Find more and updated examples in the guidelines wiki: https://osf.io/phmce/wiki/

[9] https://mybinder.org
[10] https://en.wikipedia.org/wiki/Docker_(software)
[11] https://codemeta.github.io/
[12] For example, https://travis-ci.org/, https://circleci.com/
[13] For example based on GNU make, snakemake, rake, or drake
[14] For example MIT, Apache, GPL. If you start from scratch see: https://choosealicense.com/
[15] https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1005510#sec005
[16] http://r-pkgs.had.co.nz
[17] https://github.com/crstn/UncertD3/ and http://doi.org/10.1007/978-3-319-78208-9_19

REVIEWER GUIDELINES

Reviewers of AGILE papers should be aware of the **author guidelines on reproducibility** and at the very least know the **pre-submission reproducibility checklist**, as well as the expected content of the **mandatory data and software availability section.** Using this information, reviewers should evaluate the plausibility and completeness of the data and software availability documentation, and whenever possible **include feedback on reproducibility aspects** in their comments. Reviewers should be supportive and consider potential limitations to authors due to the double-blind peer-review process.

Data and software availability documentation provide an additional set of information for assessing the quality of research presented in a contribution. Reviewers are asked to use the information from the guidelines in their assessment, and assume the position of someone who would like to reproduce the submitted work to assess whether the provided material is likely to allow reproduction of the submitted work. Reviewers are not required to actually reproduce a manuscript under review, but, if the data and code are provided in an anonymous format, and if a reviewer attempts to reproduce all or parts of the submitted work, then they are asked to document the process and outcomes.

The reviewers' comments provided to the authors are expected to be supportive and contribute to improved reproducibility of the reported findings. Reviewers can challenge authors regarding the level of reproducibility reached. Since the provision of information to help reproduction of a submitted manuscript can accidentally lead to disclosure of an author's identity, the reviewers should not use any such additional information to the disadvantage of the authors.