Website: https://osf.io/phmce/
Version: December 2020
DOI: 10.17605/OSF.IO/CB7Z8

AGILE Association of Geographic Information Laboratories in Europe

REPRODUCIBLE PAPER GUIDELINES

Full and short papers submitted to the AGILE conference **have** to include a **Data and Software Availability** section which documents data, software, and computational infrastructure to support reproduction, or mentions reasons for not publishing them.

The above requirement is the only one to comply with the AGILE Reproducible Paper Guidelines. The remainder of the document provides concrete recommendations for all involved stakeholders to increase transparency, reproducibility, and openness of computational GIScience research. The following table of contents shows the recommended parts for different readers. Familiarity with all sections is, of course, beneficial.

Further resources

These guidelines can not cover all details of the reproducibility review at AGILE conferences. For more information for authors, translations, and practical examples see the guidelines wiki. For more information about the review process and deadlines, see the process description. For any questions, please visit the AGILE Discourse server's forum for the Reproducible Paper Guidelines.

CC-BY 4.0

REPRODUCIBILITY CHECKLIST

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

- ❏ Is provided in a non-proprietary format
- ❏ Is documented for third parties to reuse
- ❏ Is accessible in a public repository and has an open data licence

For all **software tools/libraries/packages** and **computational workflows** included/produced, check if:

- ❏ Reproduction steps are explained in a README (plain text file), flowchart, or script
- ❏ Computational environments (including hardware) are documented or provided
- ❏ Versions of relevant software components (libraries, packages) are provided
- ❏ All parameters and expected execution times for the computational workflow are provided
- ❏ Software developed by the authors is available in a public repository and has an open licence
- ❏ There is a clear connection between **tables, figures, maps, and statistical values** and the data and code that they are based on, e.g., using file names or documentation in the README

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

- ❏ Data and software statements (see examples below)
- ❏ The reasons, if any, for not being able to share (parts of) data or code

For all **data and software** check that:

- ❏ All datasets and code (used or mentioned) are assigned DOIs
- ❏ Datasets and code are cited throughout the paper

After acceptance in the **camera-ready paper** check that:

- ❏ If data has been shared privately or anonymously for peer review, they are updated with all metadata and accessible via a DOI and referenced from the paper
- ❏ If a reproducibility review report will be published for your paper, a DOI URL in the Data and Software Availability section is included using the following template:
  A reproducibility report for this paper is available confirming that [considerable parts of the computational workflow / all results / Figures 1 and 4] could be independently reproduced, see https://doi.org/link_to_report.

You will find more checklists, some of them much more extensive than this one, online. If you like this style of ensuring your research is reproducible, take a look at the checklist for the *Geoscience Paper of the Future*[1] for one related to GIScience, the comprehensive *ASCE checklist*[2] with an engineering perspective, or look for checklists suited to your methods, e.g., for machine learning[3] and sharing code for machine learning[4]. The FAIR[5] (Findable, Accessible, Interoperable and Reusable) principles for scientific data management and stewardship provide further extensive guidance on improving the way we share digital assets of research.

[1] https://doi.org/10.1002/2015EA000136
[2] https://doi.org/10.1061/(ASCE)WR.1943-5452.0001215
[3] https://ai.facebook.com/blog/how-the-ai-community-can-get-serious-about-reproducibility/
[4] https://medium.com/paperswithcode/ml-code-completeness-checklist-e9127b168501
[5] https://www.go-fair.org/fair-principles/ and https://doi.org/10.5281/zenodo.2248199

AUTHOR GUIDELINES

These author guidelines will help you write the Data and Software Availability (DASA) section for your paper. They will help you to achieve basic reproducibility for data and computational workflows and learn about the extra steps you can take towards excellence in reproducibility, transparency, and openness. If the topic of reproducible computational research (RCR) is new to you, we recommend The Turing Way's Guide for Reproducible Research[6] as an accessible handbook with links to useful resources.

WRITING THE DATA AND SOFTWARE AVAILABILITY SECTION

The DASA section provides references to where data, software and documentation is available (e.g., paper section or README file) and under what conditions (e.g., copyright, licenses or access procedures for protected data). It should be concise and contain persistent links to repositories using Digital Object Identifiers[7] (DOI). You may remove links for anonymity during peer review ("xxx"), or share anonymized links[8] if your repository supports them. Data, software and (third-party) tools should be cited following recommended citation or standard citation guidelines. Possible statements for the DASA section are provided below. You may include one of these statements or draft your own.

Statements for non-computational or conceptual work

No data or code was collected, developed, or used in this work.

The full list of reviewed literature is available at [link to attachment or citable deposit of bibliography].

The full concept maps are available at [link] *and the ideas were first sketched in a blog post at* [link].

Research data/code supporting this publication …

… is available in [name of the repository(-ies)] *and is accessible via the following DOI* [DOI link(s)]

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

… was downloaded manually using the services at [name of organisation] *(using a departmental subscription for costs) and* [name of organisation]*. The compiled dataset cannot be redistributed due to licensing restrictions.*

...is not available due to [indicate reasons, e.g., licenses, sensitive data on human subjects, privacy statements; if there are processes to obtain the data, describe them].

The computational workflow supporting this publication …

… is executed via [choose, e.g., a single command/file, a workflow management software, a set of numbered scripts] *published under license* [the license] *at* [DOI of repository].

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

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

[6] https://the-turing-way.netlify.app/reproducible-research/reproducible-research.html
[7] https://en.wikipedia.org/wiki/Digital_object_identifier

INCLUDING DATA IN RESEARCH PAPERS

	Minimum requirements	Recommended practices
What?	● All input data and configuration ● Data description/documentation, including provenance, field or column types, etc. ● If data is retrieved from an external source, documentation on collection queries and download steps	● Standardised, discipline-specific metadata[8] and ontologies to describe your data ● Data download scripts
Where?	● Publish data in a public repository providing a DOI ● Cite data (including date and version) in the paper	● Discipline- or data type-specific repository[9] ● Include recommended citation in dataset description (unless already provided by repository) ● Create a registration for OSF projects[10] and use the DOI to cite it
How?	● Use open data formats; export from proprietary format for publication ● Specify the license	● Use plain text-based file formats

What if...

- **the datasets are openly available?** Cite the dataset[11]and clearly indicate which subset (if any) has been used.
- **the dataset is not openly available, is only temporarily available or is difficult to recreate?** Upload the dataset into a public repository if the original dataset license permits.
- **the licence or privacy considerations do not permit public re-sharing of the (part of) dataset?** Document the dataset and explain the procedures and conditions needed to access it. Provide a synthetic dataset to demonstrate your workflow and ideally a script for downloading.
- **you are the creator of the dataset?**: Select a license that allows the maximum reuse.
- **your data is published under your name in a public repository?** You can use **anonymised links**[12] to support anonymous review; mention the date and version of the record in the text.

Examples

- **Social media data**: If the platform's terms of service do not allow for sharing all the data in a repository provide unique identifiers of the posts used[13].
- **OpenStreetMap data**: 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, a 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, provide metadata and scraping script with all its parameters.

[8] Metadata standards catalogue: https://rdamsc.bath.ac.uk/

[9] Registry of research data repositories: https://www.re3data.org/

[10] https://help.osf.io/hc/en-us/articles/360019930893-Register-Your-Project

[11] Data Citation Principles: https://www.force11.org/datacitationprinciples

[12] Guidance how to create anonymous dataset for peer review on OSF, Figshare, Dataverse and Zenodo: https://www.cambridge.org/core/blog/2019/08/19/how-to-make-the-data-and-code-for-your-manuscript-available-to-peer-reviewers-before-making-it-public/

[13] Report on preserving social media data: https://www.dpconline.org/docs/technology-watch-reports/1486-twr16-01/file

INCLUDING COMPUTATIONAL WORKFLOWS IN RESEARCH PAPERS

	Minimum requirements	Recommended practices
What? Computational environment	• Describe the used environment and computational infrastructure, e.g., hardware specs, operating system • List software versions • Cite used software[14]	• Provide the actual environment, e.g., a Dockerfile + container[15] or a Virtual Machine (e.g., using OSGeo-Live) • Provide a pinned freeze of your dependencies (structured configuration files with dependency information) • Add a colophon or “reproducibility receipt”[16] to your notebooks • Installation and execution instructions for different operating systems
Computation steps	• Document the detailed steps in a text file and/or flowchart (every action/click) • Document expected execution times given computing power unless negligible • Ask a colleague to try out the instructions	• Scripts/models and a README file that explains their use • All figures are fully scripted and a peer has read your README’s instructions (incl. interactive visualisations and interactive adjustments • Multi-panel plots are composited with scripts[17] • Software package with structured metadata[18], tests/CI[19], and a pipeline framework[20] or workflow language[21] • Live documents for analyses, e.g., Binder[22] • Live demo of APIs/online applications (e.g., anonymous cloud resources, such as Google Cloud Run or AWS) • Subset or a synthetic dataset for quick evaluation
Where?	• Repository providing a persistent identifier, e.g., a DOI or SWHID[23]	• Versioned code repository, such as GitHub or GitLab, and ongoing open development
How? Tools used	• Use generally available tools (avoid proprietary tools that are not available to reviewers and other researchers)	• Use and create Open Source tools • Cite core modules/tools/language used
Development practices	• Use clear licenses[24] that fit your environment • Follow one of “Good enough practices in scientific computing”[25]	• Follow all “Good enough practices..” Use development guidelines for your environment / language of choice (e.g., for R[26])

Find examples for live documents, repositories with workflows and containerised computational environments in the guidelines wiki[27].

[14] Software Citation Checklist for Authors, https://doi.org/10.5281/zenodo.3479198
[15] https://doi.org/10.1371/journal.pcbi.1008316
[16] https://twitter.com/MilesMcBain/status/1263272935197782016?s=09
[17] For example, in R with patchwork, cowplot, or ggannotate, or in Python with pylustrator or Kaleido
[18] https://codemeta.github.io/ or https://citation-file-format.github.io/
[19] For example, https://travis-ci.org/, https://circleci.com/
[20] For example, GNU make, Snakemake, Nextflow, rake, drake, or targetopia
[21] For example Common Workflow Language or Workflow Description Language
[22] https://mybinder.org
[23] https://docs.softwareheritage.org/devel/swh-model/persistent-identifiers.html
[24] For example MIT, Apache 2.0, or GPL. If you start from scratch see: https://choosealicense.com/
[25] https://doi.org/10.1371/journal.pcbi.1005510
[26] http://r-pkgs.had.co.nz
[27] https://osf.io/phmce/wiki/home/

SCIENTIFIC REVIEWER GUIDELINES

This section clarifies the expectations and role of the scientific reviewer with respect to the reproducible paper guidelines. For information for the Reproducibility Reviewer, please see the following section.

> Reproducibility is considered good scientific practice that provides input for the quality assessment of a paper. Therefore, reviewers of AGILE papers should be aware of the **author guidelines on reproducibility** and be familiar with the **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 and readily available **include feedback on reproducibility aspects** in their comments. Scientific reviewers are free to but **are not expected to attempt reproductions of computations**.

Data and software availability documentation provide an additional set of information for assessing the quality of research presented in a manuscript. Reviewers are asked to know about the AGILE reproducible paper guidelines and to consider the level of reproducibility reached in a manuscript. To do so, they shall 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. Based on this impression, reviewers may challenge authors regarding the level of reproducibility reached, if any statements are made regarding reproducibility in a manuscript.

Scientific reviewers are not required to actually reproduce a manuscript, 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 (see Reproducibility Reviewer Guidelines below). Please reach out to the reproducibility chair if you are keen on conducting a reproducibility review for a paper you are reviewing.

The peer review of AGILE papers is a fully anonymous peer review, i.e. authors and reviewers do not know each other's identity. Reviewers should be supportive to authors and consider potential limitations in access to resources due to anonymisation. Since the provision of information to help reproduction of a paper 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. The reviewers' comments provided to the authors are expected to be neutral[28] and contribute to improved reproducibility of the reported findings.

[28] https://doi.org/10.1038/d41586-020-03394-y

REPRODUCIBILITY REVIEWER GUIDELINES

Reproducibility reviewers conduct a complimentary review of the computational workflow that is published with a full paper that is provisionally accepted after the scientific review process. They read the paper insofar as needed to **reproduce the computation**, **using the abstract and the Data and Software Availability section** (DASA) as starting points. Ideally, these sections of the paper together with a README file are sufficient for the reproduction. When reproducibility reviewers get stuck, they take advantage of the option to **communicate** with the authors early and often. Reproducibility reviewers should be aware of the different reproducibility levels (see Author Guidelines above) to **recommend improvements** to the authors, but they are not responsible for making a workflow transparent or executable. Reproducibility reviewers **write a reproducibility report** documenting the results of their reproduction attempt and their communication with the authors. The report is published if the reproduction was, at least in part, successful. It is shared with the authors if the reproduction attempt was stopped but already contains relevant feedback.

Reproducibility review coordination

The reproducibility chair will be your contact person regarding supporting infrastructure and getting access to the private discussion forum for reproducibility reviewers on the AGILE Discourse server[29]. This forum is used to assign, under the leadership of the reproducibility chair, the reproducibility reviewers to papers matching their respective topical and technical skills, and share material such as a current template for the reproducibility report.

Goals and scope

While the AGILE reproducible paper guidelines are created with the intention to eventually have 100% reproducibility success rate for accepted papers, the road to this goal is positive encouragement, understanding, and ultimately community adoption through conviction. This makes a clear definition of your tasks as reproducibility reviewer harder and progress slower yet hopefully more sustainable. A reproducibility review is an extra merit for an accepted paper, but a successful reproduction is not a requirement for acceptance. The reproducibility reviewer should be aware of this supporting role and accept that not all authors might "take the extra few steps" needed. This non-exclusionary practice is also reflected by the fact that only one reproducibility reviewer is assigned per paper. You may be both the reproducibility reviewer and the scientific reviewer on the same paper, but the roles of the two types of reviewers are complementary. The scope of the reproducibility review is roughly in line with the CODECHECK principles[30], and the CODECHECK community is worth exploring for further examples and materials for an effective reproducibility review. A *partial reproduction*, e.g., the recreation of some but not all of the figures should still be seen as a success at this point, though what is "good enough" may change over time. Please consult with your fellow reproducibility reviewers or the reproducibility committee chair in case of doubt.

Reproducibility reviewer skills

A reproducibility review is a learning experience for both authors and reviewers and part of a process for the AGILE community to increase openness and transparency. Therefore, these guidelines do not mention a fixed amount of time you should spend on a reproduction attempt, as your experience or interest are just as unique as the research you are tasked to reproduce. However, we suggest erring on the side of "not my fault" after a few minutes of being stuck and not spending more than an hour to get a workflow started. However, this depends also on your interest, time budget, and skills with the given software and libraries. We recommend to get basic familiarity with package managers and virtual environments, e.g., pip, Conda and venv for Python, DESCRIPTION files and renv for R, npm for JavaScript, and even containerisation with Docker. Use the reproducibility reviewer discussion forum early and often for your questions and issues.

[29] https://discourse.agile-online.org/
[30] https://codecheck.org.uk/#the-codecheck-principles

Do's and don'ts of a reproducibility review

Do

Quick pre-repro-review checks and ask authors to fix before continuing, e.g., using the reproducibility checklist; even if not all of these are technically required, authors who are willing to work reproducibly can show their engagement right from the start:

1. Do the links to data sets and materials resolve?
2. Is there a README with clear step-by-step instructions?
3. Is there a clear mention of to be expected execution times?
4. Is there a LICENSE file to ensure openness?

Encourage authors by pointing out promising intermediate results or concrete benefits of reproducibility.

Accept sample datasets to run a workflow and compare the outcome with the expected sample results; check the sources of the full datasets, if available.

Clearly document the extent of the reproduction in your reproduction report and suggest potential improvements; if you provide intermediate feedback, to include a history of your interactions in the report so that the ideas you contributed are preserved when the paper's material is improved.

Get in touch with fellow reproducibility reviewers if specific expertise (tool, programming language, ..) is needed.

Set an example when communicating about computational problems, e.g., by clearly defining your system (OS version, language version, etc.)

Ask specific questions or point out concrete problems that may lead authors to improve their material, including referencing these guidelines or concrete tools/methods that you already (!) know about, especially if you suspect that the author might now be familiar with them (e.g., version pinning/dependency management, absolute paths).

Make sure that you are aware of any templates or specific resources provided for reproducibility reviewers from the reproducibility committee chair before starting your review.

Consider the author's background, career stage, and position to be aware of (a lack of) privileges or institutional power to decide how much support you provide and how you communicate; your reproducibility review can be a contribution to improve equity and inclusion in academia.

Don't

Dig across badly or un-documented collections of files and functions to identify which part of the code/data creates which figure/table/output.
Attempt to find or build the "start button" yourself.

Run workflows requiring considerable computational resources (unless interesting for you) but ask for data subsets for demonstration purposes.

Accept private sharing of data or code, unless strictly required for protection of sensitive data. All changes by the author should update to the public reproduction material.

Attempt to install software without any instructions, install binary software of unknown origin, or try to fix installation problems you encounter on your machine; try to install without (a) asking for help from a fellow reproducibility reviewer who is familiar with the software, or (b) asking the author to help, providing a minimal reproducible example of your problem.

Point out or even fix problems that are not specific to the paper, e.g., general problems in a software tool.

Create new accounts on any service or platform merely to be able to access code, data, or other resources. Code, data, and other resources should be accessible on a platform without requiring registration.

Fix anything (unless you really enjoy doing so), e.g.,

- compiler problems,
- outdated libraries,
- broken paths, or
- Incomplete computing environment specifications,

especially if the author can fix them even quicker.

Be a bro[31].

[31] https://thepsychologist.bps.org.uk/volume-33/november-2020/bropenscience-broken-science

BACKGROUND

The AGILE Reproducible Paper Guidelines are part of AGILE's mission to promote research, education and networking of the geoinformatics community in Europe and beyond. Reproducibility of research is a pillar of science. It is closely connected with transparency and reusability. All these aspects are crucial for trust in science and enable better science. Despite the expected high standards of scholarly papers, textual contributions consistently fall short in allowing readers to assess the computational aspects of research work[32]. Research in geoinformatics and GIScience and related fields frequently involves data and computational methods. For these methods, **every step toward higher reproducibility counts**. Therefore, 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. Our guidelines focus on computational reproducibility: the data and software is provided in a way that the analysis may be re-run to reproduce the study results[33]. While it is also critical that experimental protocols and hardware (e.g., 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.

A primary concern is **documentation**. The level of documentation should be so good, that the computations can be completed by a reviewer without deep knowledge of the language or libraries. The level should be suitable for someone who can grasp the *abstract* of the paper, not for someone who totally understands the paper's methods section. While open access to data and code is an asset for enabling reproducibility, reasons for refraining from making (parts of) the material openly available exist. These reasons need to be disclosed. The structure and documentation (e.g., README files, file organisation and naming conventions) of data, software, and computational infrastructure should be sufficiently clear to allow readers to recreate the original authors' computational environment as closely as possible to recreate the original results. Furthermore, author should be aware of **benefits**, such as selfish reasons for reproducibility[34] or individual benefits from data citation[35]. The community can value not only **contributions** to a body of literature, but also to data, code and best practices.

The improvement of reproducibility is a **cultural change** that is intertwined with many bigger challenges in science, such as broken metrics, fair evaluation, and healthy working environments. To contribute to this change and take along the AGILE community, the guidelines have minimal requirements and include recommended practices to strive for. The primary maxim for everyone involved in creating, reviewing and interacting with papers 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.

This document further elaborates on different **roles of reviewers** in the current process to make clear that reproducible papers are only in part a technical challenge but largely a culture of conducting research. Scientific reviewers judge scientific merit and reproducibility reviewers validate outcomes of a workflow match the advertised results. Distinct roles further allow to include people with different skill sets, give flexibility with respect to the type of review, allow to engage early career researchers in peer review, and share the workload. In general, reproducibility creates opportunities for increased interaction within the community: methods can be tested and applied to further study areas, methods can be extended and improved, latest research can be used in an educational context, and in-depth feedback on research work can be provided. More concretely, reproducibility reviews can spark new collaborations through joint improvements of workflows, and are in all cases educational for both author and reproducibility reviewer.

These guidelines provide a positive setting for researchers of all career stages, skill levels, and backgrounds to face the challenges of reproducibility together. In the long run, all conference papers become reproducible and the special attention to reproducibility wanes as reproducibility becomes common for members of all AGILE member labs, because they internalise its advantages and reap its rewards with **high quality research**.

[32] https://theconversation.com/how-computers-broke-science-and-what-we-can-do-to-fix-it-49938

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

[34] https://doi.org/10.1186/s13059-015-0850-7

[35] https://doi.org/10.1371/journal.pone.0230416

CC-BY 4.0