An Architecture for Reproducible Computational Geosciences

Daniel Nüst
Institute for Geoinformatics,
University of Münster

Münster, Germany
daniel.nuest@uni-muenster.de

Marc Schutzeichel
University and State Library,
University of Münster

Münster, Germany
m.schutzeichel@uni-muenster.de

Abstract

Reproducibility is a cornerstone of science but poses a large challenge when it comes to modern computational sciences. Initiatives for Openness must be accompanied by an infrastructure going beyond the state of the art in scientific publications and preservation of knowledge. Building on the concept of Executable Research Compendia (ERC), this work presents an architecture to support a scholarly process for computational geosciences. In this architecture the novel reproducibility service enriches scientific publications and integrates with the existing platforms.

Keywords: reproducible research, computational geosciences, ERC, software architecture

1 Motivation

Reproducibility is a cornerstone of science but poses a large challenge when it comes to modern computational sciences. Data, methods and products are all digital: from inception/measurement, via algorithmic analyses to static and interactive online publications. Putting aside meta-challenges such as a community-wide definition of the term "reproducible", the triplet of Open Source Software, Open Science projects and Open Access publications has created unprecedented potential to collaborate in all steps of a scientific process: idea, implementation, scholarly review, publication, and preservation. If everything is out in the open, there should be more scrutiny of existing work, less repetition of basics, and higher degree and quality of collaboration.

According to good scientific practice, all research should be reproducible by nature. But day-to-day obstacles and the pressure of academia lead to publishing first and foremost articles and rarely complete workflows. To break the modus operandi, we see two courses of action.

On the one hand we see organisational support, e.g. incentives and accreditation promising higher visibility of reproducible works, mandatory supplemental materials by journals, updated teaching contents, and proper funding.

On the other hand technical solutions, i.e. tools and services making it easier to conduct reproducible research and to leverage the advantages of reproducible analyses.

We see the latter as a crucial point of vantage. By easing the way towards reproducible scientific publications and preserving knowledge instead of collecting citations, the geosciences community can reach new levels with respect to how reviews are being conducted and how publications can be used. This work introduces a novel technical building block, the *reproducibility service*. The challenges are not unique to computational geosciences, but we argue the solutions must be.

2 An Architecture for Reproducible Geosciences

The stakeholders involved in the process of computational geosciences, from inception of an idea to preservation of the results, are scientists (author, reviewer, editor, reader), publishers, service operators (publisher, research institution), and curators. This work provides a common viewpoint and language for them. It builds on the concept of Executable Research Compendia (ERC). ERCs are container combining text (documentation, the actual publication), code (analysis scripts, runtime environment), data, and user interface specifications (facilitating manipulation by readers) in a meaningful manner (cf. Nüst 2017).

The *reproducibility service* integrates with existing services and platforms involved in the publication and archival of geosciences research by providing the following functions:

- create ERC from provided workspaces (including semi-automatic metadata extraction and elicitation from the user), which is initiated from publication platforms
- save ERC to data repositories and archives
- execute ERC in scalable computing infrastructures using trusted data repositories
- save ERC metadata in registries to facilitate discovery

This comprises a relevant extension of the ERC's self-containment idea at the execution stage, which is crucial for geosciences.

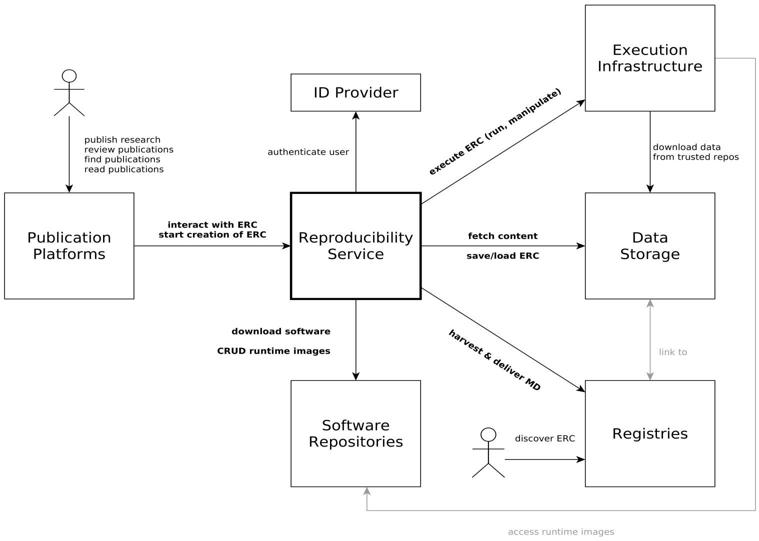
The idea of data services and APIs is well-established (cf. OGC standards, http://www.opengeospatial.org/docs/is).

Data sets in domains such as remote sensing can easily reach volumes too large for single files or disks. Therefore a selection of trusted data repositories and APIs can be defined by the reproducibility service to limit data duplication while still using ERC as transferable and archivable units.

Taking into account the scientific setting, the following aspects are crucial qualities of the reproducibility service:

- *transparency*: allow scrutiny required by a rigorous scientific process

Figure 1: Overview of architecture for reproducible geosciences.



- *integration*: connect with existing platforms and focus on the core functionality, i.e. do not replicate complex tasks such as long-term storage, established and accepted procedures such as peer-review procedure, or interdisciplinary efforts such as persistent identifiers

Figure 1 places the *reproducibility service* in the context of scholarly publications. It (a) enhances current practices in computational geosciences from publishing static documents to sharing executable, interactive publications, and (b) integrates with existing services for peer review, publication, data storage and long-term archival.

The *execution infrastructure* is domain-agnostic thanks to the abstraction provided by ERC. The *registries*, e.g. DataCite (http://datacite.org) or CrossRef (http://www.crossref.org), as well as ID providers, e.g. ORCID (http://orcid.org), must be domain independent as they serve cross-cutting goals.

The *software repositories* and *data repositories* for storage and archival of ERC are not different from any other scientific area. Examples for the former are GitHub (http://github.com) for source code, apt (https://wiki.debian.org/Apt) for operating system packages, CRAN (http://cran.r-project.org) for language-specific extension packages, or Docker Hub (http://hub.docker.com) for runtime images. Examples for the latter are Zenodo (https://zenodo.org) or figshare (https://figshare.com).

The *data repositores providing data* however are specific to geosciences domains. They must be accepted by domain members and provide the required data. Candidates here are platforms such as PANGAEA (https://www.pangaea.de), GFZ data services (http://dataservices.gfz-potsdam.de) or Copernicus Scientific Data Hub (https://scihub.copernicus.eu).

The *publication platforms* are the author's, reviewer's, and reader's main contact point with ERC and naturally essential. They, too, are domain-specific and require awareness of challenges and of the importance in reproducibility of computational analysis within the specific community.

3 Summary & Outlook

The architecture presented here is a work in progress report on software and concepts developed in the national research project Opening ReproducibleResearch (http://o2r.info) funded by the German Research Foundation (DFG) under project no. PE 1632/10-1, KR 3930/3-1 and TR 864/6-1). It identifies domain-agnostic conceptual components and points out those functions specific to the geosciences, namely data storage platforms & data service access during execution of ERC.

Following the spirit of Open Science, the architecture is developed publicly in a repository on GitHub (https://github.com/o2r-project/architecture). The authors welcome external contributions. Suggestions and improvements by the geospatial community will improve the ongoing development of the reproducibility service as well as supporting tools and demonstrators.

References

Nüst, Daniel, Markus Konkol, Edzer Pebesma, Christian Kray, Marc Schutzeichel, Holger Przibytzin, and Jörg Lorenz. 2017. "Opening the Publication Process with Executable Research Compendia." *D-Lib Magazine* 23 (January). 10.1045/january2017-nuest.