

An Architecture for Reproducible Computational Geosciences

MOTIVATION & THE REPRODUCIBILITY SERVICE

Data, methods and products of geoscience research today are digital: from inception/measurement, via algorithmic analyses to static and interactive online publications. 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.

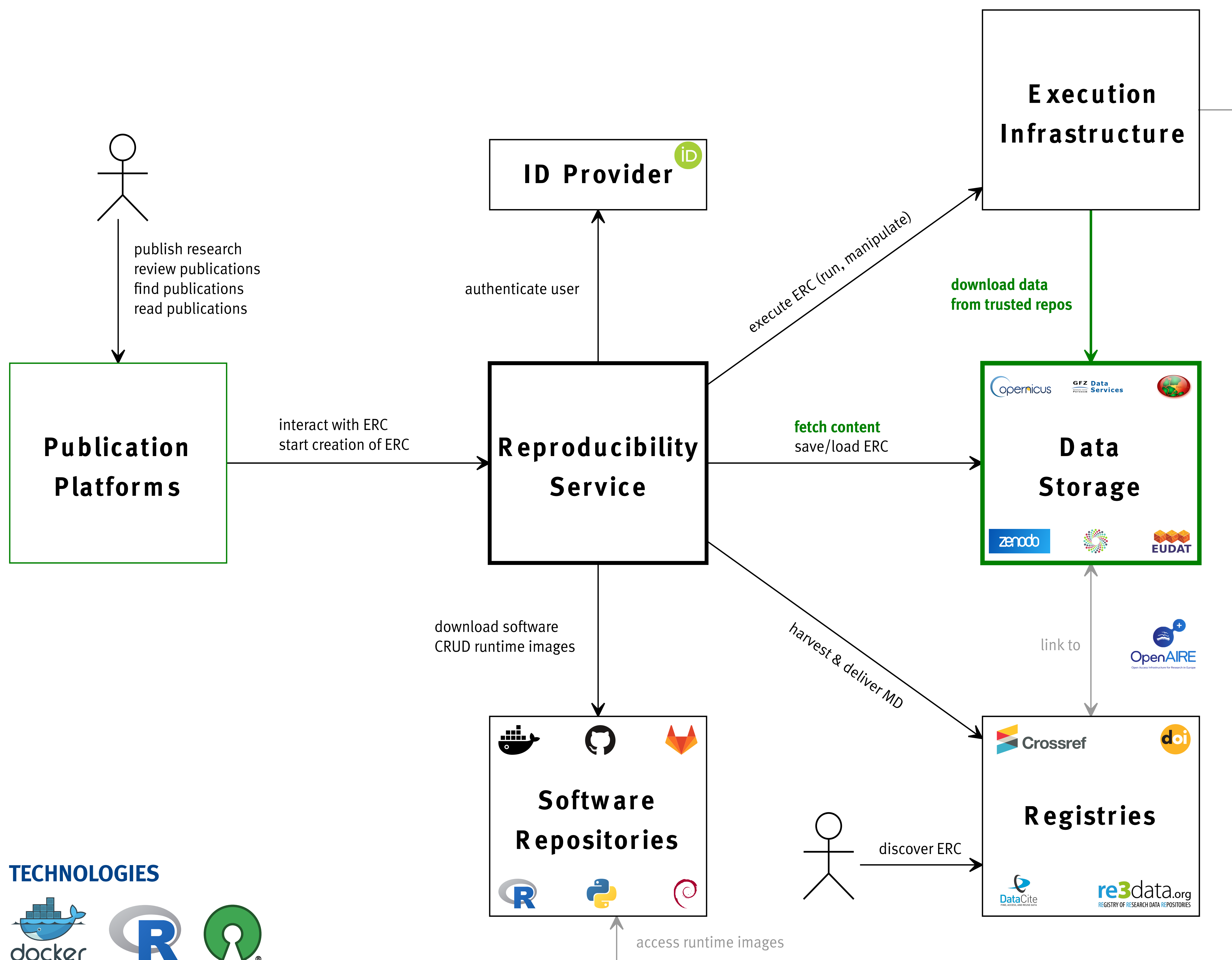
There should be more scrutiny of existing work, less repetition of basics, and higher degree and quality of collaboration, but the *pressure of academia* leads to publishing first and foremost articles and rarely complete workflows.

To break the modus operandi, we see *supporting technical solutions*, i.e. tools and services making it easier to conduct reproducible research and to leverage the advantages of reproducible analyses, as a crucial point towards reproducible scientific publications. The must be accompanied by improvements in education, scholarly accreditation mechanisms, and **scientific culture**.

By **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.

Two of its qualities are crucial in the scientific setting: **transparency** allows scrutiny required by a rigorous scientific process; **integration** with existing platforms ensures quality through focus on the core functionality and easier adoption.

The service does not replicate the complex tasks of data storage, peer-review procedures, preservation, or interdisciplinary efforts such as persistent identifiers. It enhances current practices in computational geosciences from publishing static documents to sharing executable research compendia.



GEO-READY

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 initiated from publication platforms
- save ERC to data repositories and archives
- execute ERC in scalable computing infrastructures **allowing connections** to 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 **data repositories** are also specific to geoscience domains, because they must be accepted by domain members and provide the required data.

PROJECT

o2r opening reproducible research

<http://o2r.info>

@o2r_project

Opening Reproducible Research is a joint project by the Institute for Geoinformatics and the University and Regional Library Münster funded by the German Research Foundation aims to improve the exchange and archiving of geoscience research results that are published over the Internet, by facilitating access to them and by simplifying their reuse in the form of a research compendium. The project focuses on the publication process and the interaction with digital research objects as technical challenges to increase publication of open reproducible research. Combining these goals will set a clear innovation in the area of reproducibility, which is at the heart of the scientific method.

SUMMARY

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.

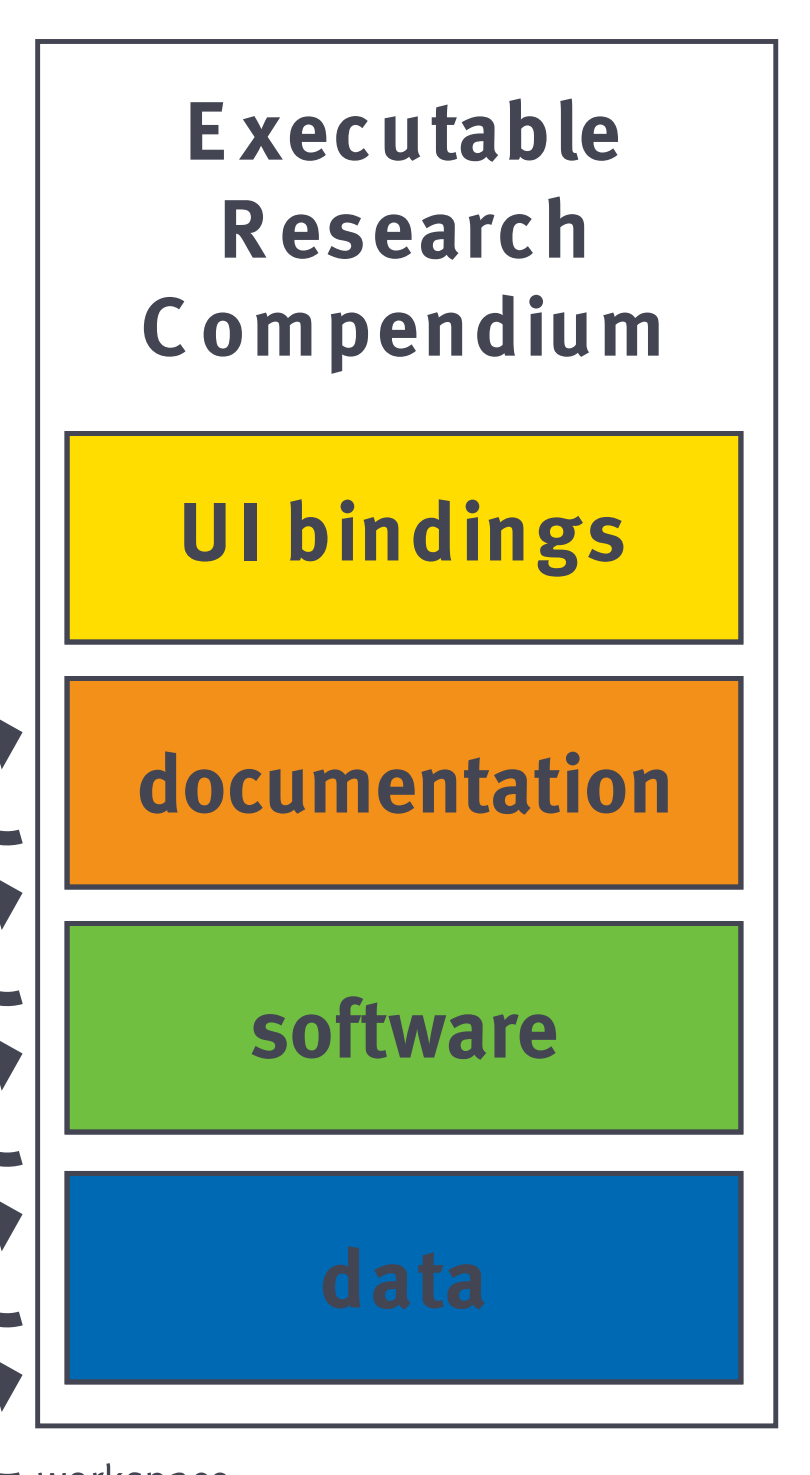
The architecture presented here is a work in progress report on software and concepts. It identifies domain-agnostic conceptual components and points out those **functions specific to the geosciences**, namely data storage platforms & data service access during execution of ERC. Following the spirit of Open Science, the architecture is developed publicly in a repository on GitHub. Suggestions and improvements by the geospatial community are welcome.

Contribute at <https://github.com/o2r-project/architecture>

BACKGROUND: ERC

Executable Research Compendia (ERC) support requirements of authors, readers, publishers, curators, as well as preservationists. They are a new way to package computational research combining data, software, text, and a user interface description and provide a novel potential to find, explore, reuse, and archive computer-based research. [1]

Data comprises all inputs for an analysis, ideally starting with raw measurements, in form of text files, or databases. **Software** comprises analysis code/scripts created by a researcher and the complete runtime environment as an executable Docker image and a Dockerfile as the manifest. **Documentation** comprises both instructions (e.g. a README), the actual scientific publication, and metadata in standardized formats (licenses, discovery metadata). The actual publication comes in a source format (i.e. based on literate programming) and a viewable format (e.g. an HTML document). **UI bindings** open up the compendium. They allow reviewers to interact with diagrams and manipulate formerly hidden parameters for a comprehensive understanding of the underlying data and code. A **formal specification** for ERC connects these building blocks in a meaningful way. It enables technical checking of computation outputs of an ERC and closes the gap of dependency preservation for computational scholarly works.



[1] Opening the Publication Process with Executable Research Compendia
Nüst, D., Konkol, M. et al.
D-Lib Magazine, 2017
doi: 10.1045/january2017-nuest