

Conquaire: Continuous quality control for research data to ensure reproducibility

Vidya Ayer, Philipp Cimiano, Fabian Herrmann, Vitali Peil, Christian Pietsch, Andreas Rempel, Jochen Schirwagen, Johanna Vompras, Cord Wiljes
contact: christian.pietsch@uni-bielefeld.de – ORCID: <https://orcid.org/0000-0001-8778-1273> – license: CC-BY 4.0 – last modified: March 15, 2019
<https://conquaire.uni-bielefeld.de>

About

We present progress in **Conquaire**, a DFG project to **foster the analytical reproducibility of research results** at Bielefeld University, Germany, conducted by CITEC (Center of Excellence in Cognitive Information Technology), Bielefeld University Library and 9 research groups from 2016 to 2019. DFG grant number: 277747081.

Motivation

Reproducibility of research results enables peer validation. However, reproducing research results is a major challenge:

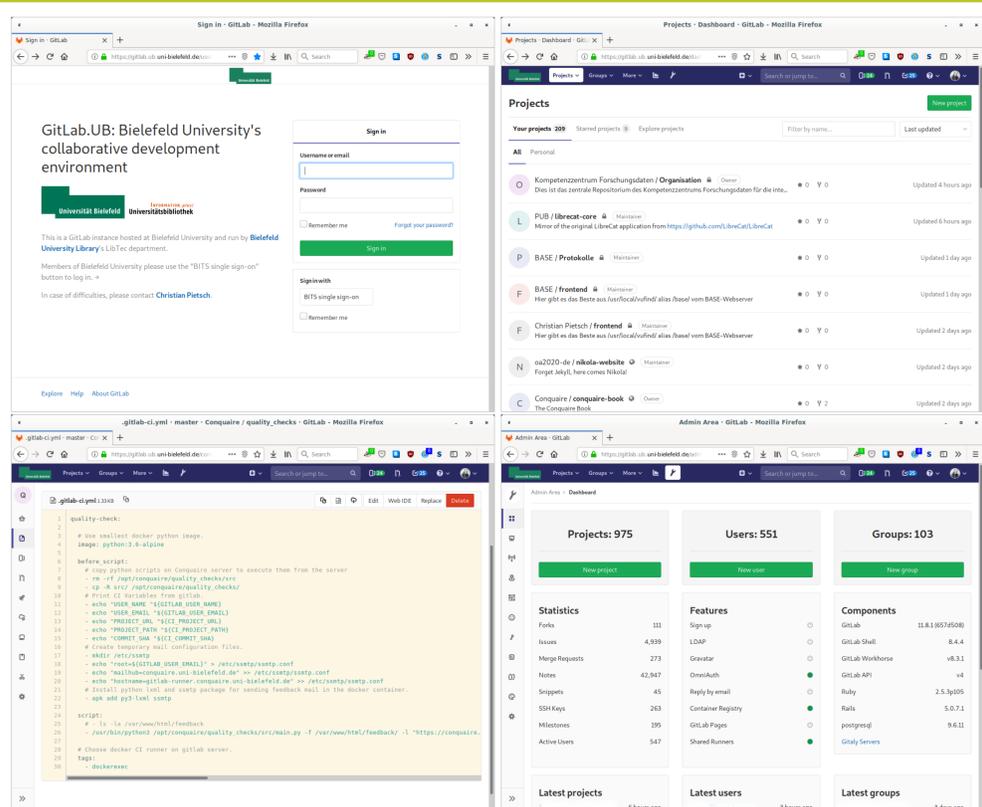
- Nature survey (1576 scientists): **50–70 % reproducibility failures**
- Reproducibility Project: **61 % failure rate in Psychology**
- pharmaceutical clinical drug trials: **82 % failure rate**

Conquaire approach to improve data quality

Prime directive: Do not create any extra effort for researchers, and not to stand in their way. Instead, we try to tap into their existing workflows in **4 easy steps:**

1. offer an institutional **GitLab** instance with Large File Storage
2. offer **training** in **Git** and **GitLab** as well as for **PUB** (see below)
3. offer **data quality checks** implemented in **GitLab CI** (see below)
4. connect **GitLab** to our institutional publication repository **PUB** for instant data publishing incl. **DOI** minting, **linking**, and **archiving**

1. GitLab.UB



2. Training

- In 2016, delegates from 8 research groups (project partners) attended several hands-on Git training sessions based on the **Software Carpentry** course <https://swcarpentry.github.io/git-novice/>.
- Once in each semester since summer 2018, we have offered a Git and GitLab training session to Bielefeld University staff. They were almost fully booked.

3. Continuous data quality checks

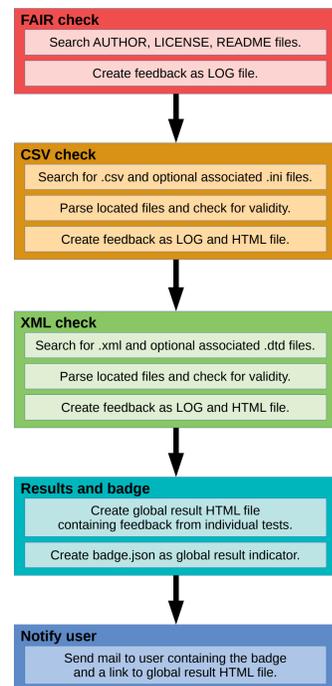
Inspired by successful software development practices such as **versioning** and **continuous integration (CI)**, we apply these ideas to research data management (RDM).

The **GitLab** component **GitLab CI** can automatically run arbitrary programs whenever files are added or changed. These programs run on a separate server within a temporary **Docker** container.

To set this up for a **GitLab** project, all that a user has to do is to put a file called **.gitlab-ci.yml** in its main directory.

We provide such a file pre-configured to run a Python script we developed.

It checks for basic **FAIR data principles** compliance and tries to validate the most popular data file types.



4. Publishing datasets

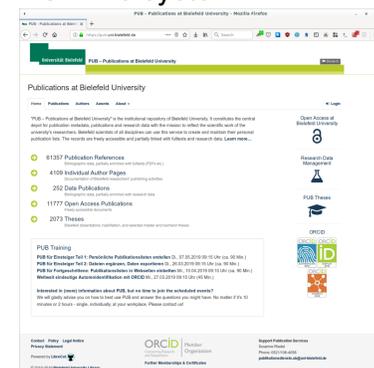
Bielefeld University's institutional publication repository **PUB** is the place where researchers enter their publications, including data publications. **PUB** is an in-house development based on the free and open-source **LibreCat** software.

As an alternative to the existing upload button, researchers can now select one of their **GitLab.UB** projects to **publish any version** and get a **DOI** for it. **PUB** talks to the **GitLab API**.

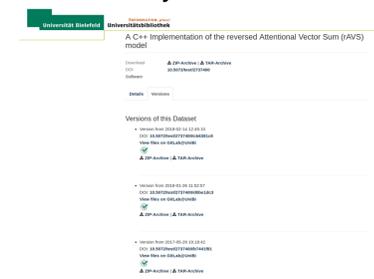
Data publications can be linked to conventional publications. Metadata can be re-used in the **Linked Open Data** spirit via an open API in formats such as **MODS** and **DataCite XML**.

Soon, **PUB** will display a **badge** indicating the overall result of the data quality checks from step 3.

PUB live system:



PUB test system:



Related work

This project was initiated by Najko Jahn's observation that researchers started to use **GitHub** for tasks other than software development (see the project proposal on the **Conquaire** website for quantitative details).

Later, **GitHub** started to co-operate with **Zenodo** to offer a data publication service not unlike the one we provide for Bielefeld University but without pre-configured quality checks.

More recently, **GitLab** added a feature called **Auto DevOps** that provides pre-configured quality and security checks for software development.

Workshop invitation

You are cordially invited to the **Conquaire Workshop on Data Quality and Reproducibility** taking place in Bielefeld's ZIF on 3 April 2019. Please register soon: <https://www.uni-bielefeld.de/conquaire/reproducibility-workshop/>