Universität Bielefeld

INFORMATION.plus!
Universitätsbibliothek

CITEC
Cognitive Interaction Technology
Cluster of Excellence
Bielefeld University

Conquaire

Continuous quality control for research data to ensure reproducibility ⟶ http://conquaire.uni-bielefeld.de | Poster License: CC BY-NC-ND 4.0

Vidya Ayer, Christian Pietsch, Johanna Vompras, Jochen Schirrwagen, Cord Wiljes, Vitali Peil & Philipp Cimiano {Contact: ayer, cpietsch}@uni-bielefeld.de

Introduction

We present Conquaire, a DFG-funded project to foster the reproducibility of research results at Bielefeld University, Germany.

- a generic infrastructure to enable analytical reproducibility,
- it builds upon GitLab, a web interface for the version control system Git, to support data sharing and access to different versions of research data
- adopts continuous integration principles to improve data quality, enabling reuse and reproducibility of the published analytical results

Motivation

- Reproducibility of scientific research: Essential principle of science.
- Peer-verified by research community.
- Reproducing research results is a major challenge:
  - A Nature survey among 1576 scientists found 50–70 % reproducibility failures.
  - In Psychology, the Reproducibility Project reported a 39 % success rate.
  - Pharmaceutical clinical drug trials have a lower success rate of 18 %.

Goals: ‘non zero-sum research’

- Create a generic research data management system (RDMS) to manage the data and scripts for publications.
- Computational reproducibility of a statistical analysis is mathematical, hence verifiable.
- Adopt continuous quality control principles for research data to ensure reproducibility, data reuse and data sharing.
- Meet open data and open research standards: data artifacts (primary, secondary and original programs) used by the researcher are openly accessible.
- Data quality management: syntactic validity, semantic integrity tests, etc.

Project partners

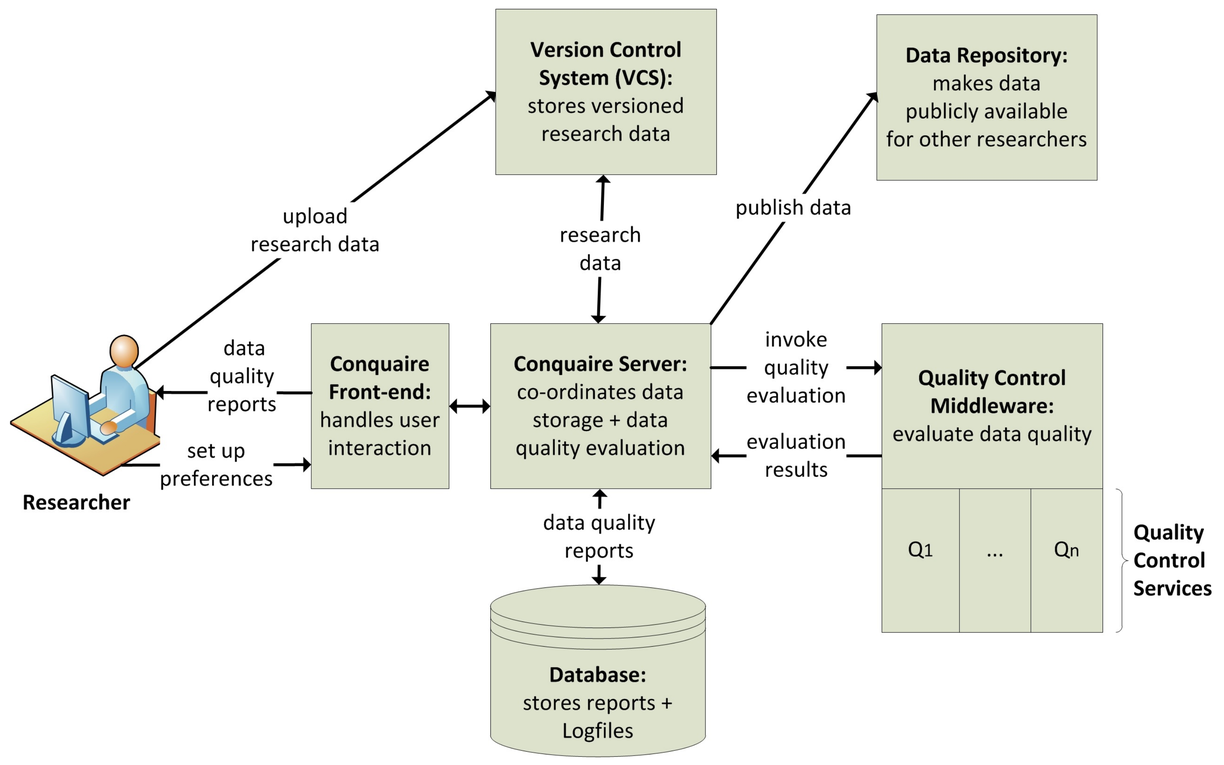
- Disciplines: Biology, Computer Science, Applied Computational Linguistics, Neurobiology, Sports Science, Neurocognitive Psychology, Atmospheric and Physical Chemistry, Economics and Linguistics.
- Software: Python, Pandas, R, C, C++, Matlab, SPSS.
- Data Storage: MySQL, Dropbox, Sciebo, private servers, backup drives, HPC cluster, etc..
- Data Formats: CSV, XLS, XML, JSON, JPEG, MP4, EAF (ELAN annotated files).

GitLab based research workflow

- values of analysed data that have changed between two (or more) commits
- keep track of user interactions over time
- ability to revert or experiment with data over the research project time frame
- timestamps and version control provides a timeline for the data, results and program scripts

Conquaire Architecture

- GUI Frontend: Provides a visual interaction interface enabling researcher to upload/view data, visualise quality reports, tag specific releases, etc..
- VCS Server: GitLab server supports versioned storage using a non-linear development workflow allows multiple users to branch/ merge research data objects.
- Conquaire Server: Message-oriented generic infrastructure framework monitoring and controlling research data artifacts committed into GitLab.
- Quality-Control Middleware: Performs quality checks, ensures analytical/computational standards, and provides data feedback.
- Conquaire Database: Only store metadata information and data tags in a semantic data format.
- Trusted Data Repository: Releases are published in PUB (Bielefeld University’s institutional repository) and assigned a DOI via DataCite. PUB is archived in the SAFE Private LOCKSS Network.



First Quality Experiment: Development Status

Our first minimal proof-of-concept development for Quality is under testing and uses Python for development:

- CSV file checks: Quality checks the CSV file for out-of-range numeric values, NAN / Null values, data types.
- CI Runners: GitLab’s continuous integration (CI) runners are used to monitor commits made into Gitlab.
- Flask: Receives Git push events via GitLab’s web hook API.