

ConQQuaire

Quality Checks

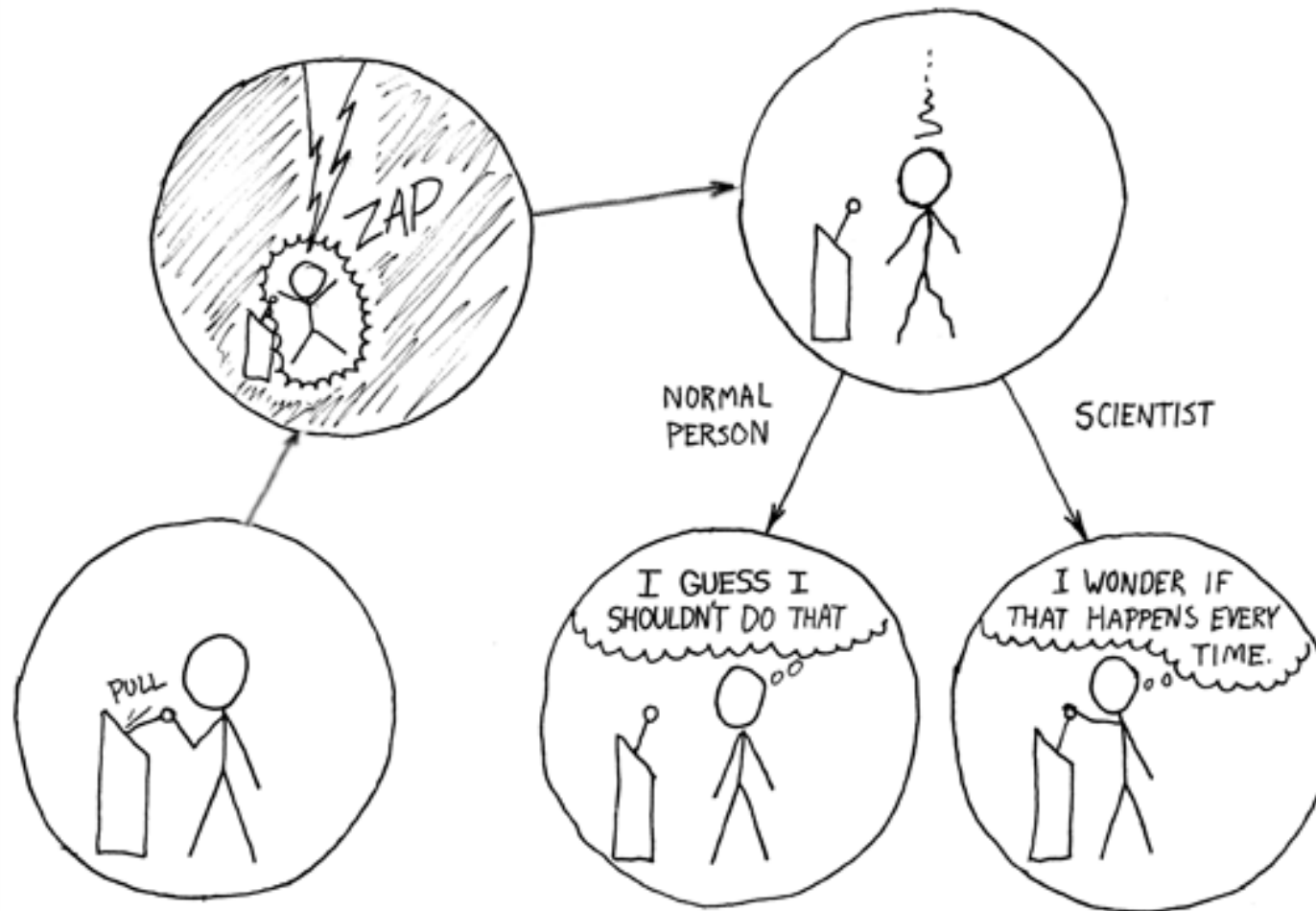
Fabian Herrmann

fherrmann@techfak.uni-bielefeld.de

Table of Contents

- Introduction to quality checks
- Quality checks workflow in Conquaire
 1. FAIR check
 2. Filetype specific checks
 3. Results and badge
- How to use Conquaire quality checks
- Requirements for Conquaire quality checks
- Benefits for data publications
- Summary

Reproducibility



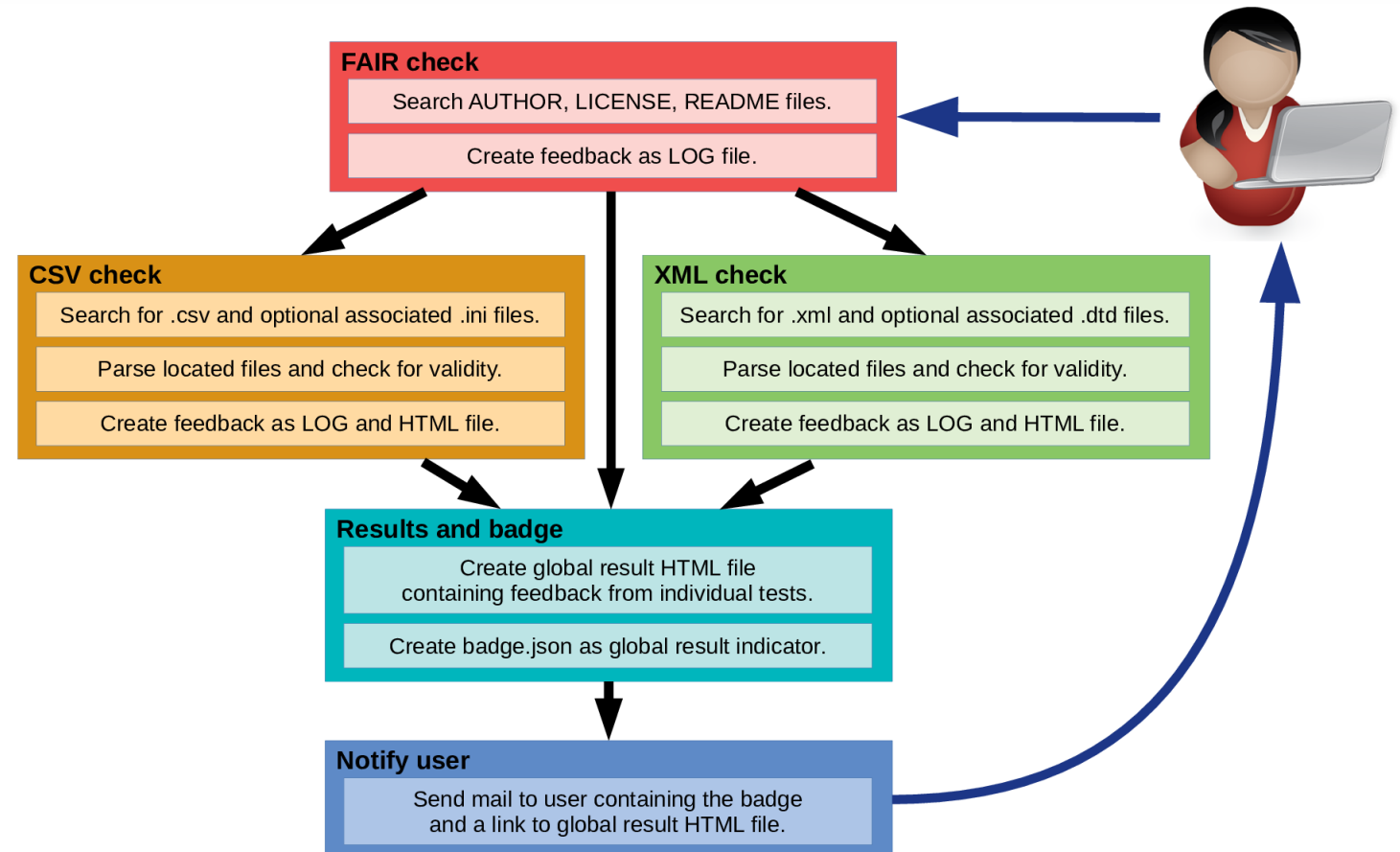
source: xkcd.com

Idea Behind Our Implementation of Quality Checks

- improve data quality without interfering with current research practices
 - automatic background process
- ensure that scientific data is in a usable state
 - guarantee valid data for further research
- make use of common standards
 - coverage of CSV and XML files
- quality checks are written in Python 3.6 to provide open source standard

Workflow Overview

- **User** adds data to repository and **commits**
- automatic workflow
 1. FAIR check
 2. Filetype specific checks
 3. Results and badge
- **User is informed** about results via email



1. FAIR Check

- adaptive implementation of the FAIR checks (<http://fairmetrics.org/>)
- check for essential files
 - AUTHOR (Who has created the data?)
 - LICENSE (How can the data be used for further research?)
 - README (What is the data about?)

FAIR metrics

- ✓ /quality_checks/AUTHORS.md [LOG](#)
- ✗ /quality_checks/LICENSE [LOG](#)
- ✓ /quality_checks/README.md [LOG](#)

2. CSV/XML/... Checks

CSV (comma separated file)

- usage of Python CSV library
- check for well-formed files
 - accessible
 - header, consistent number of rows/columns
- provide optional .ini file to check for additional data criteria:
 - column type, range of values

	Ozone	Solar.R	Wind	Temp	Month	Day
1	41	190	7.4	67	5	1
2	36	118	8	72	5	2
3	12	149	12.6	74	5	3
4	18	313	11.5	62	5	4
5	NA	NA	14.3	56	5	5
6	28	NA	14.9	66	5	6
7	23	299	8.6	65	5	7
8	19	99	13.8	59	5	8
9	8	19	20.1	61	5	9
10	NA	194	8.6	69	5	10

2. CSV/XML/... Checks

XML (extensible markup language)

- use of external LXML library
- check for well-formed file
 - accessible
 - correct syntax of tags
- provide optional .dtd file to check if data complies with schema declarations

```
1 <?xml version="1.0" encoding="UTF8" ?>
2 <node_description>
3     <target id="windows 64bit">
4         <graphics>nvidia_870</graphics>
5         <power_plug_type>energenie_eu</power_plug_type>
6         <tes>performance test</tes>
7     </targetREREREREREREt>
8 </node_description>
```


3. Results and Badge

FAIR metrics

- ✓ /quality_checks/AUTHORS.md [LOG](#)
- ✗ /quality_checks/LICENSE [LOG](#)
- ✓ /quality_checks/README.md [LOG](#)

CSV checks

- ⚠ /quality_checks/data/csv_data/airquality-xpt-2018mar29.csv [LOG](#) [HTML](#)
- ⚠ /quality_checks/data/csv_data/airquality.csv [LOG](#) [HTML](#)
- ⚠ /quality_checks/data/csv_data/airquality2.csv [LOG](#) [HTML](#)
- ⚠ /quality_checks/data/csv_data/rdm-course_survey_results.csv [LOG](#) [HTML](#)

XML checks

- ⚠ /quality_checks/data/xml_data/airquality.xml [LOG](#) [HTML](#)
- ✓ /quality_checks/data/xml_data/book_db.xml [LOG](#) [HTML](#)
- ✗ /quality_checks/data/xml_data/book_db2.xml [LOG](#) [HTML](#)
- ⚠ /quality_checks/data/xml_data/rdm-course_survey_results.xml [LOG](#) [HTML](#)
- ✗ /quality_checks/data/xml_data/sample.xml [LOG](#) [HTML](#)

Quelle:

https://conquaire.uni-bielefeld.de/feedback/fherrmann/quality_checks/be039664af6c5baa8255f23b2b7f1c1e055037cc/result.html

- three possible badges

- valid



- well-formed

- not well-formed



- easy to understand

- also accessible for colour-blind people



- overall badge = worst individual check result

Content of Feedback Mail

Von root <root@gitlab-runner.conquaire.uni-bielefeld.de> ☆ ↶ Antworten → Weiterleiten 📁 Archivieren 🔥 Junk 🗑️ Löschen Mehr ▾

Betreff **Quality check results for repository https://gitlab.ub.uni-bielefeld.de/fherrmann/quality_checks** 13.02.19, 17:37

An f.herrmann@uni-bielefeld.de ☆

Hello Fabian Herrmann,

you've been running some tests on the repository https://gitlab.ub.uni-bielefeld.de/fherrmann/quality_checks with commit id 5a872046...

This is the overall result that would be displayed in PUB if you publish your data in the current state:

✖ Your data is not well-formed.

The detailed test results are stored under following link: https://conquaire.uni-bielefeld.de/feedback/fherrmann/quality_checks/.../result.html

In the summary, for every file that was tested, an icon indicates the test results.

Further information is provided as LOG files and optional HTML files which are linked.

This mail was autogenerated. Please do not respond.

How to Use Conquaire Quality Checks?

1. Place preconfigured `.gitlab-ci.yml` in your repository directory
2. Add file to your Git (*`git add .gitlab-ci.yml`*)
3. Add a commit (*`git commit -am "Your commit message."`*)
4. Push changes to your Git (*`git push`*)
5. You will receive a mail with a link to your feedback.

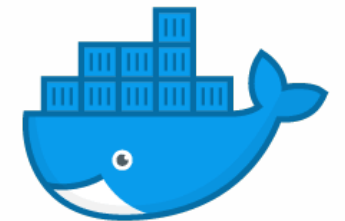
Requirements for Conquaire Quality Checks

- Server with GitLab instance including a CI runner
- Installation of docker
 - Python 3.6 alpine image
 - Within container: installation of external libraries LXML and sSMTP
 - is provided by preconfigured YAML file
- Installation of mail service
- Accessible web server



pythonTM

source: wikipedia.org



docker

source: docker.com


→ www.uni-bielefeld.de/citec






Benefits for Data Publications

- GitLab is connected to PUB University of Bielefeld
- create data publication directly from GitLab
 - no need of double upload
- PUB fetches feedback results automatically (submitted via JSON)
 - public badge indicates data quality

PUB LIKATIONEN
UNIVERSITÄT
BIELEFELD

Research Data Management Course: Survey Data

Wiljes C (2018)
Bielefeld University. 

Download  LICENSE.txt 20.44 KB
 README.txt 2.15 KB
 rdm-
course_survey_questions.pdf
165.14 KB
 rdm-course_survey_results.csv
7.44 KB
 rdm-course_survey_results.xlsx
11.77 KB
— Less

DOI 10.4119/unibi/2920783
Datenpublikation

Details Dateien Links Zitieren

Summary

- create opportunity to ensure reproducibility of research results
- easy handling for all users
- low requirements to use Conquaire quality checks
- simple integration of further filetype specific tests
- improvement of publication standard for data publications in PUB
University of Bielefeld

ConQuaire

Thank you very much!

Fabian Herrmann

fherrmann@techfak.uni-bielefeld.de