UNIVERSITÄT BIELEFELD

Tag der Forschungsdaten in NRW 2021

Kontinuierliche Qualitätskontrolle von Forschungsdaten zur Sicherung der Reproduzierbarkeit von Forschungsergebnissen

Prof. Dr. Philipp Cimiano
Universität Bielefeld

16.11.2021

UNIVERSITÄT BIELEFELD

DFG-Projekt Conquaire

„Continuous quality control for research data to ensure reproducibility”

Ziele:

1. Verbesserung der Qualität von Forschungsdaten
2. Stärkung der analytischen Reproduzierbarkeit von Forschungsergebnissen

UNIVERSITÄT BIELEFELD

Ziel 1: Verbesserung der Qualität von Forschungsdaten

UNIVERSITÄT BIELEFELD

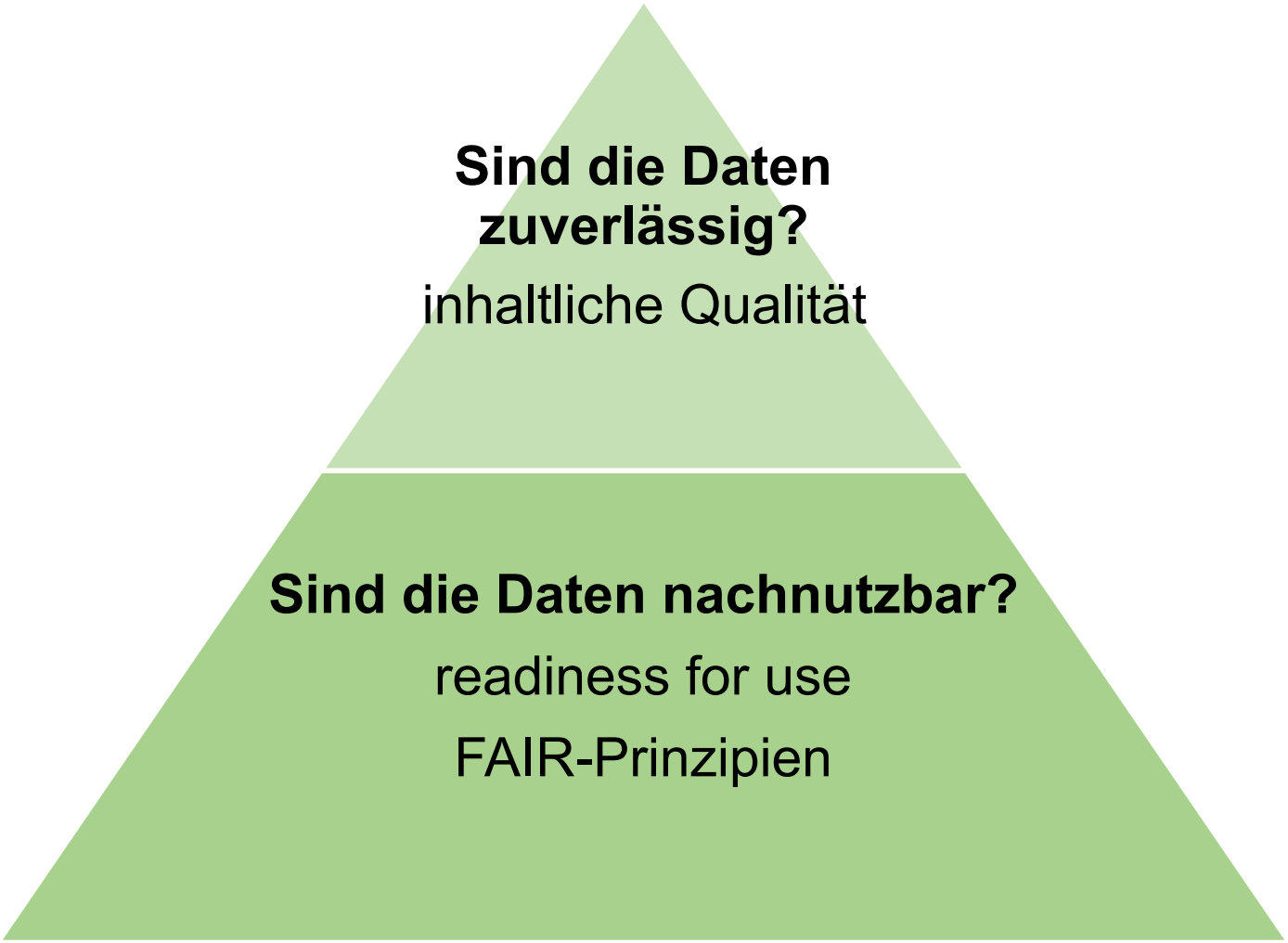
RfII: Herausforderung Datenqualität

„Gefordert ist ein gemeinsamer **Qualitätsdiskurs** aller Akteure unter der Voraussetzung einer genauen Analyse der Veränderungen, die der digitale Wandel für methodische Forschung mit sich bringt, wie auch verbindlicher **Qualitätssicherungsmaßnahmen** in der Wissenschaft.“

Rat für Informationsinfrastrukturen (RfII): Herausforderung Datenqualität – Empfehlungen zur Zukunftsfähigkeit von Forschung im digitalen Wandel, Göttingen 2019.

UNIVERSITÄT BIELEFELD

Was ist „Datenqualität“?



UNIVERSITÄT BIELEFELD

Was ist Datenqualität?

- Nachnutzbarkeit = Readiness for (re-)use
- FAIR-Prinzipien:
  - **F**indable - auffindbar
  - **A**ccessible - zugänglich
  - **I**nteroperable - interoperabel
  - **R**e-usable - wiederverwendbar

UNIVERSITÄT BIELEFELD

Idee von Conquaire

- Unterstützung für Forschende durch forschungsbegleitende automatisierte Überprüfung der Datenqualität
- Realisierung auf Basis von „Continuous Integration“

UNIVERSITÄT BIELEFELD

Continuous Integration (CI)

- Technologie aus der Softewareentwicklung
- Ziel: Steigerung der Softwarequalität
- Beim Einstellen von Dateien in die Versionsverwaltung wird automatisch eine Aktion durchgeführt
- Typische Aktionen:
  - Übersetzen und Linken der Anwendungsteile
  - automatisierte Funktions-Tests
  - Softwaremetriken zur Messung der Softwarequalität
  - …

UNIVERSITÄT BIELEFELD

Versionierung mit Git

- Versionskontrollsystem für Programmcode und textbasierte Dateien
- revisionssichere Historie der Änderungen
- ältere Versionen können jederzeit wiederhergestellt werden
- lückenloser Nachweis, wer wann welche Änderungen durchgeführt hat

UNIVERSITÄT BIELEFELD

GitLab

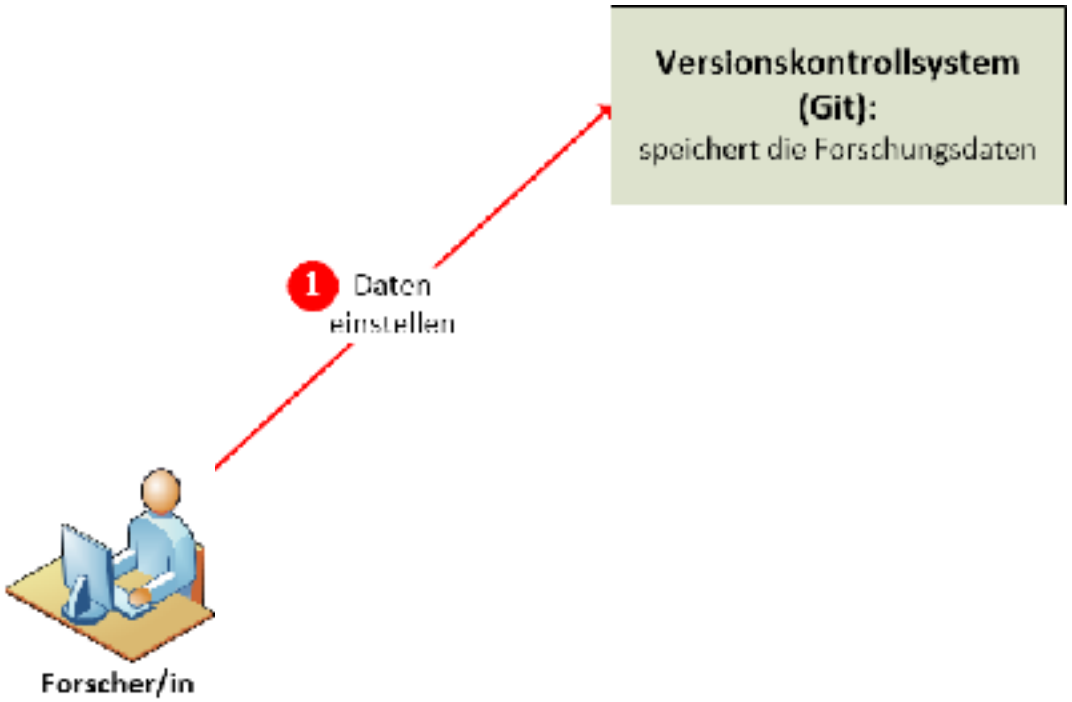
- User-Interface für Git
- Zusätzlich:
  - Projektmanagement-Features
  - Entwicklungstools
  - GitLab CI

UNIVERSITÄT BIELEFELD

Conquaire Systemarchitektur

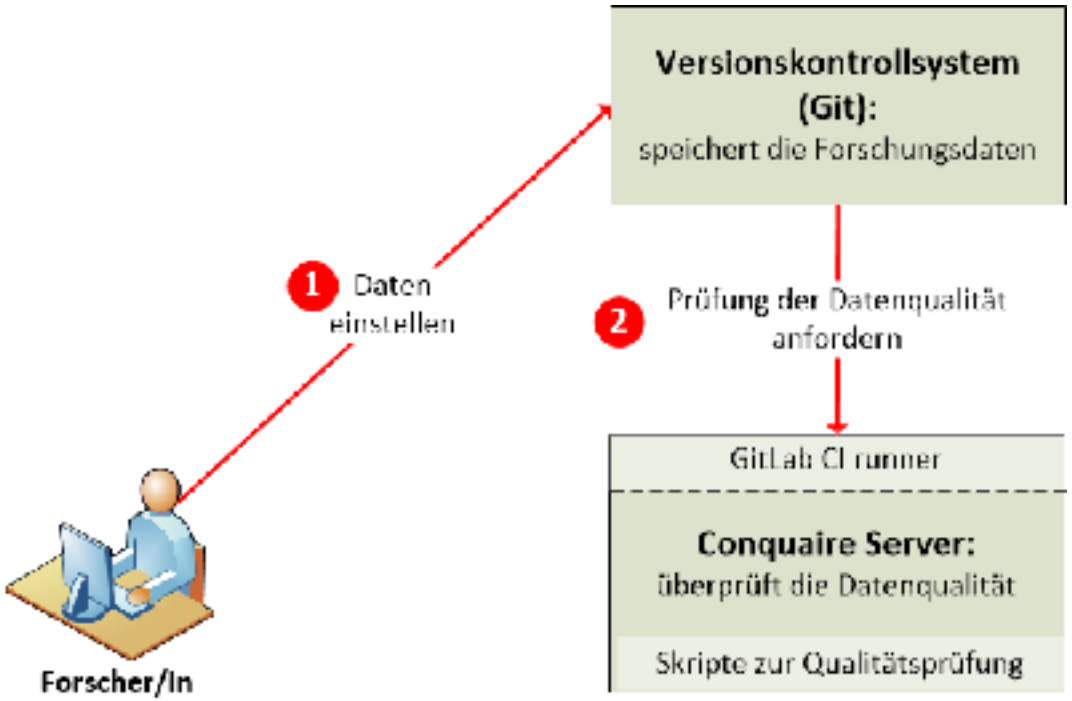
UNIVERSITÄT BIELEFELD

Conquaire Systemarchitektur



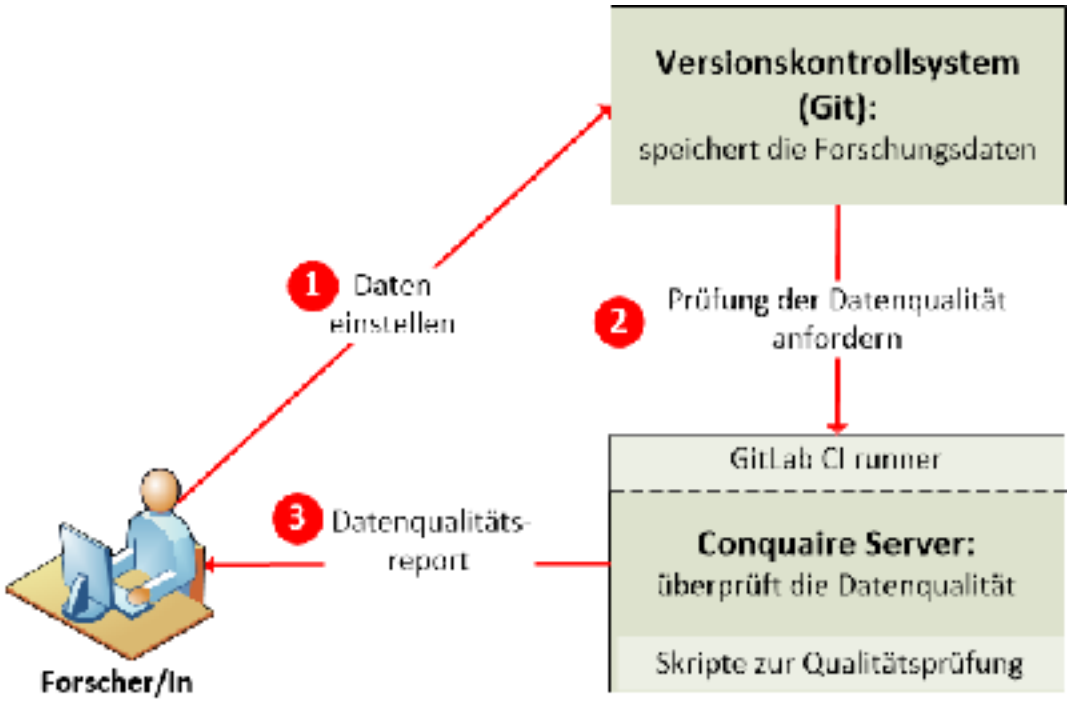
UNIVERSITÄT BIELEFELD

Conquaire Systemarchitektur



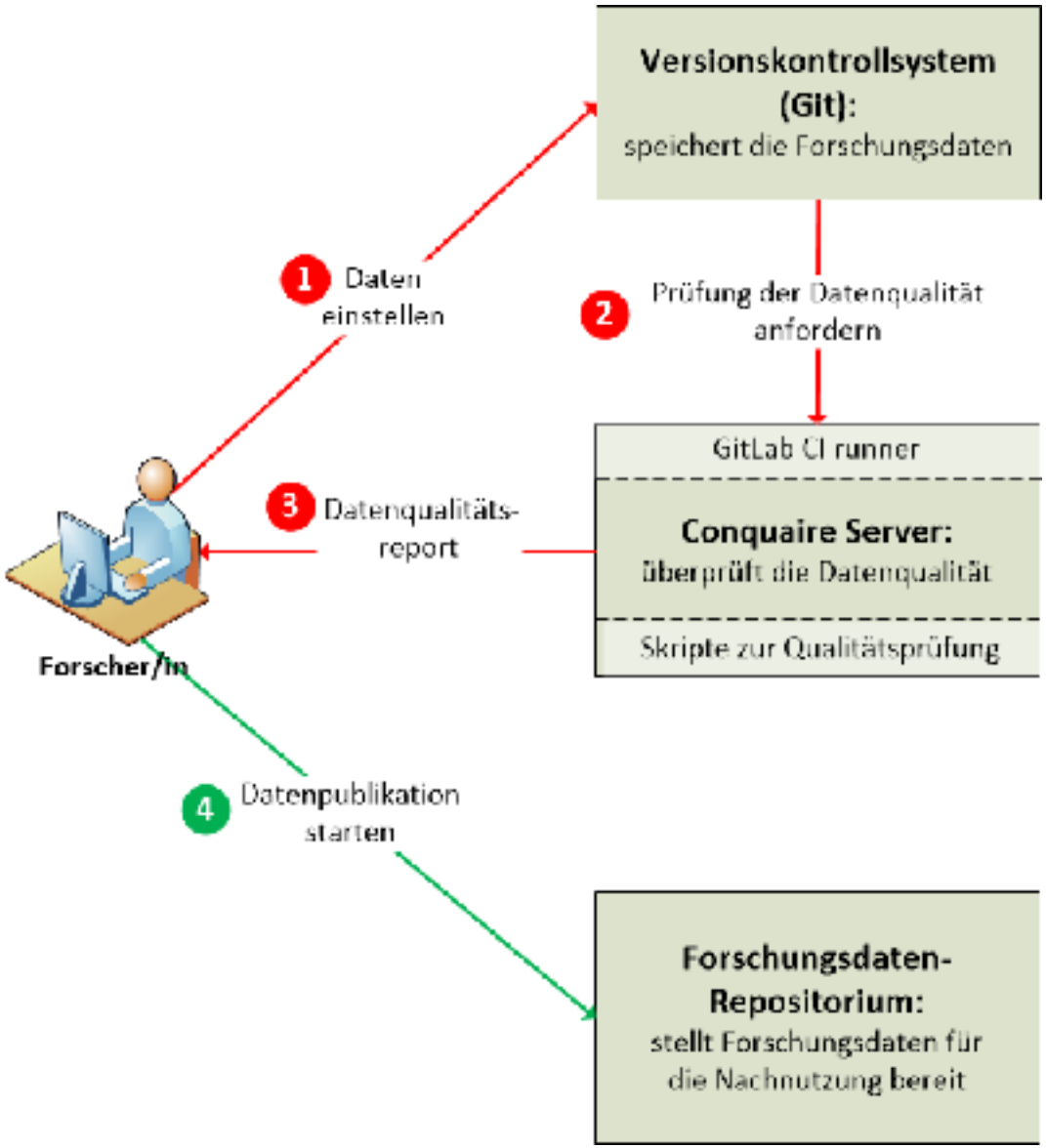
UNIVERSITÄT BIELEFELD

Conquaire Systemarchitektur



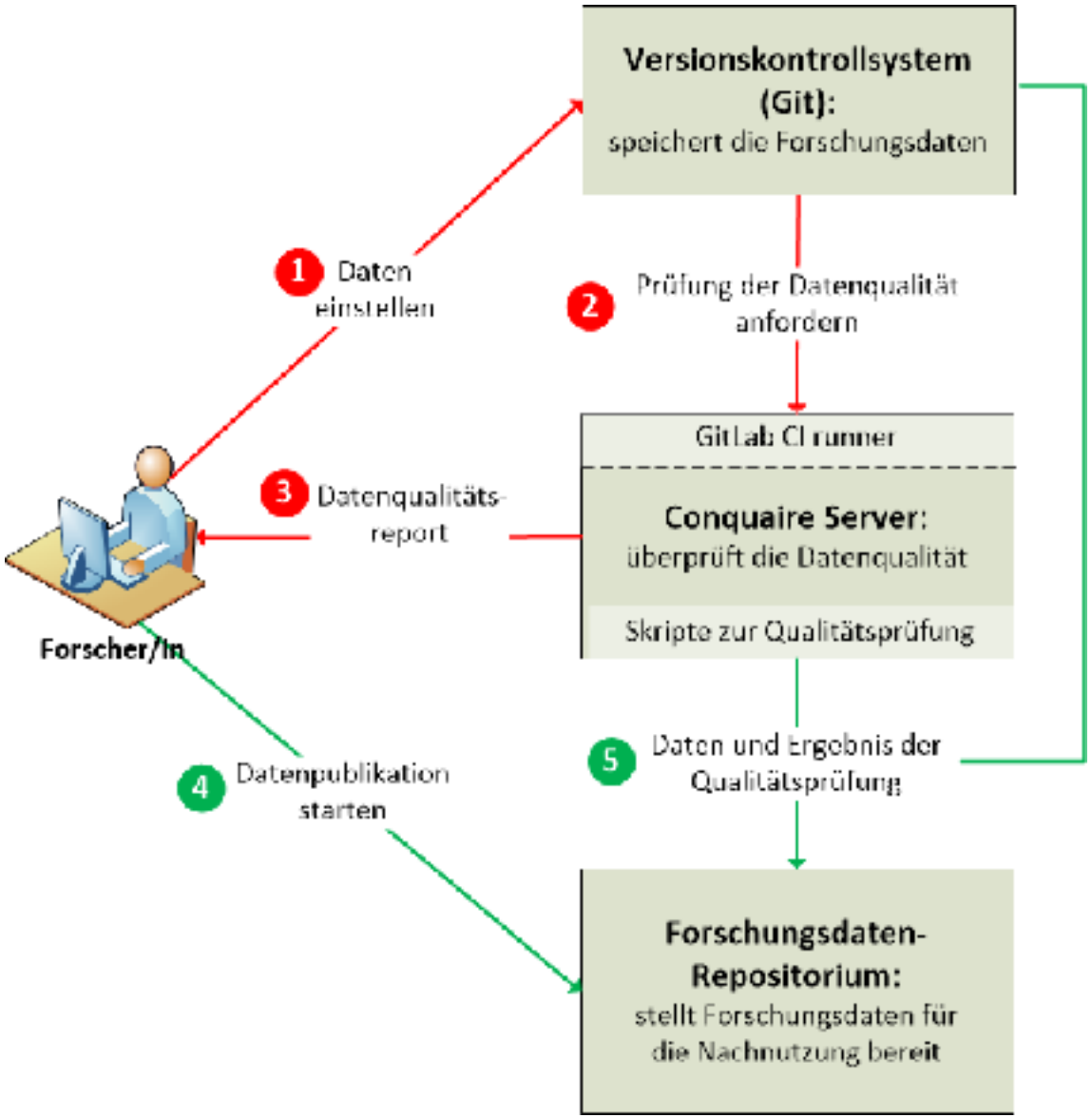
UNIVERSITÄT BIELEFELD

Conquaire Systemarchitektur



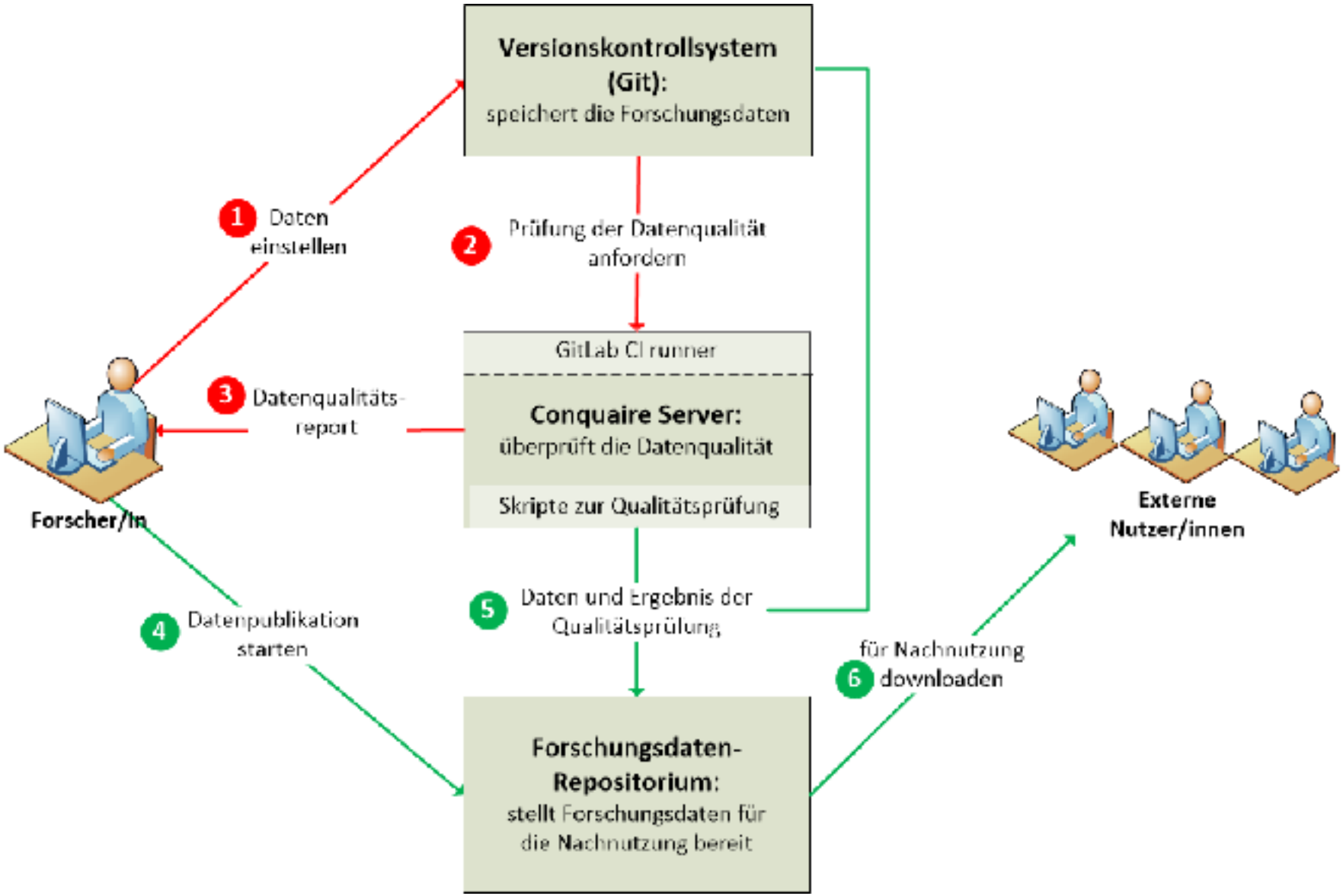
UNIVERSITÄT BIELEFELD

Conquaire Systemarchitektur



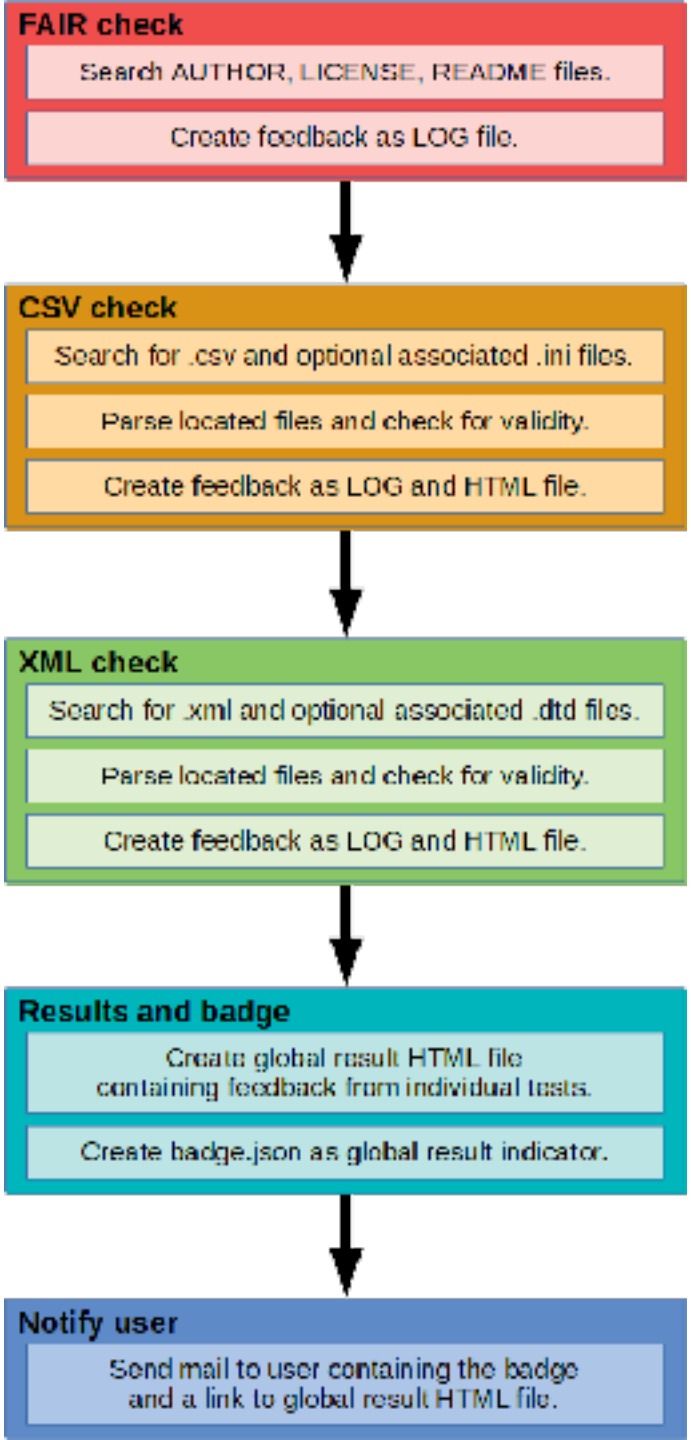
UNIVERSITÄT BIELEFELD

Conquaire Systemarchitektur



UNIVERSITÄT BIELEFELD

Conquaire Pipeline



UNIVERSITÄT BIELEFELD

Qulitäts-Badges

O.K. (=syntaktisch korrekt und valide)

Warnung

Fehler

UNIVERSITÄT BIELEFELD

Qualitätsbericht via E-Mail

FAIR metrics

✅ /quality_checks/AUTHORS.md LOG
❌ /quality_checks/LICENSE.md 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

UNIVERSITÄT BIELEFELD

Qualitätsbericht: CSV

/quality_checks/data/csv_data/airquality2.csv

```
[ERR] in row 1, column "Ozone": not an integer
[WRN] in row 5, column "Ozone": undefined entry
[WRN] in row 5, column "Solar.R": undefined entry
[WRN] in row 6, column "Solar.R": undefined entry
[WRN] in row 10, column "Ozone": undefined entry
[WRN] in row 11, column "Solar.R": undefined entry
[WRN] in row 25, column "Ozone": undefined entry
[WRN] in row 26, column "Ozone": undefined entry
[WRN] in row 27, column "Ozone": undefined entry
[WRN] in row 27, column "Solar.R": undefined entry
[WRN] in row 29, column "Month": value too large
```

	Ozone	Solar.R	Wind	Temp	Month	Day
1	ERROR	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

UNIVERSITÄT BIELEFELD

Qualitätsbericht: XML

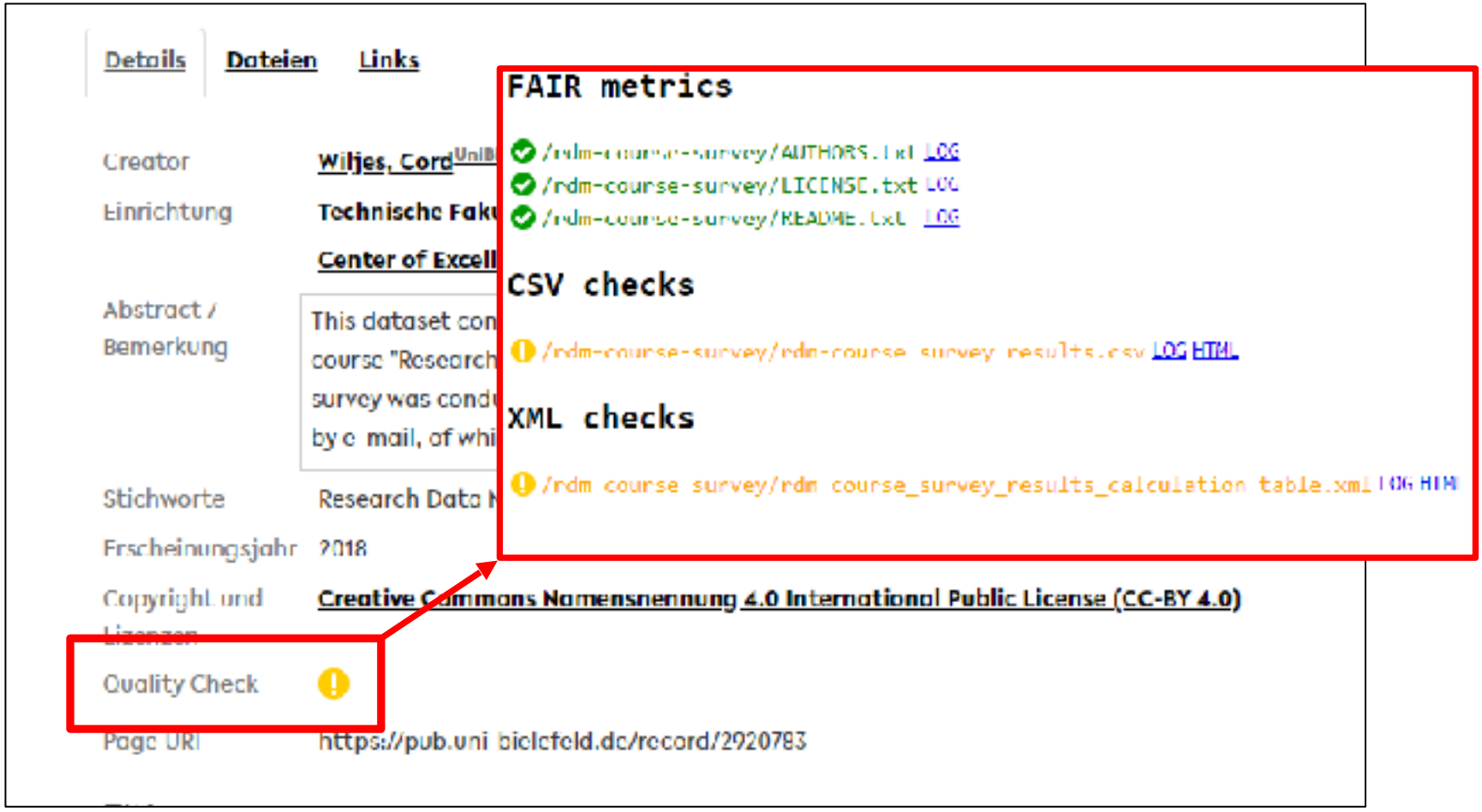
/quality_checks/data/xml_data/sample.xml

```
[WRN] in line 8: Element target content does not follow the DTD
[WRN] in line 11: No declaration for element tes
```

```
 1 <?xml version="1.0" encoding="UTF8" ?>
 2 <node_description>
 3     <target id="windows 32bit">
 4         <graphics>nvidia_970</graphics>
 5         <power_plug_type>energenie_eu</power_plug_type>
 6         <test>unit test</test>
 7     </target>
 8     <target id="windows 64bit">
 9        <graphics>nvidia_870</graphics>
10        <power_plug_type>energenie_eu</power_plug_type>
11        <tes>performance test</tes>
12     </target>
13 </node_description>
```

UNIVERSITÄT BIELEFELD

Integration in institutionelles Repositorium



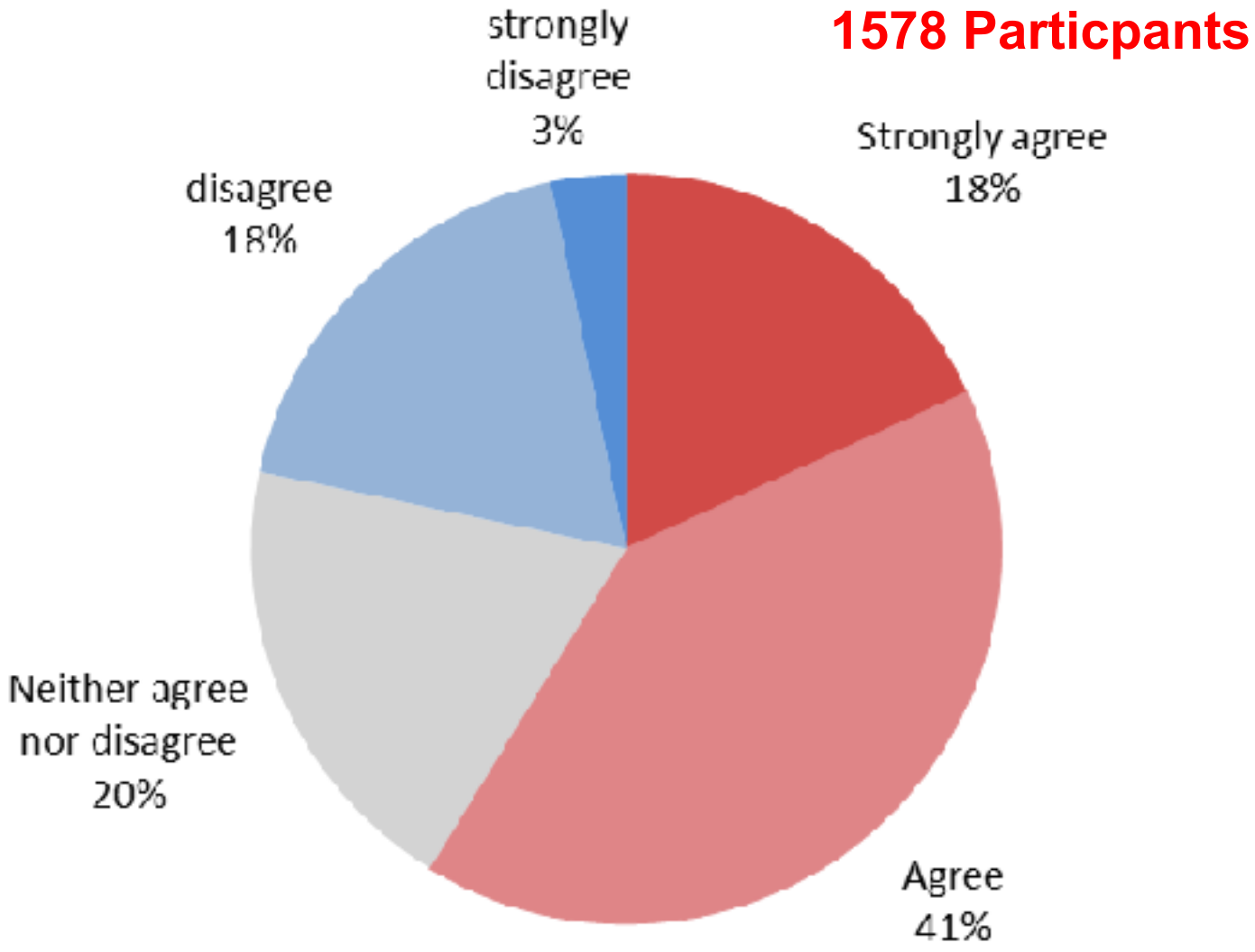
Wiljes, C. (2018). *Research Data Management Course: Survey Data*. Bielefeld University. https://doi.org/10.4119/unibi/2920783

UNIVERSITÄT BIELEFELD

Ziel 2: Stärkung der analytischen Reproduzierbarkeit von Forschungsergebnissen

"I think that the failure to reproduce scientific studies is a major problem in my field"

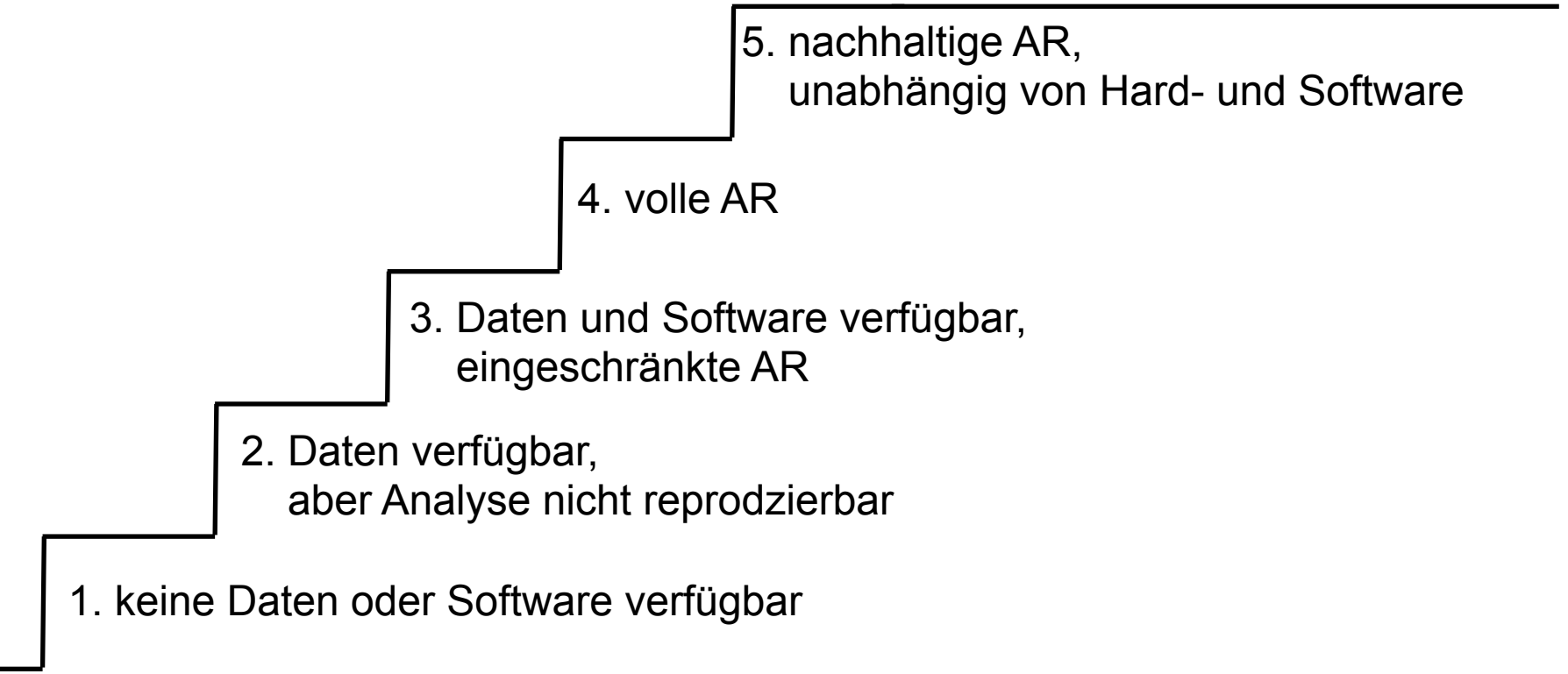
1578 Particpants



Nature 533, 452–454 (26 May 2016)

UNIVERSITÄT BIELEFELD

Analytische Reproduzierbarkeit
= Reproduzierbarkeit der Datenanalyse



UNIVERSITÄT BIELEFELD

Partnerprojekte

Prof. Martin Egelhaaf
Biologie

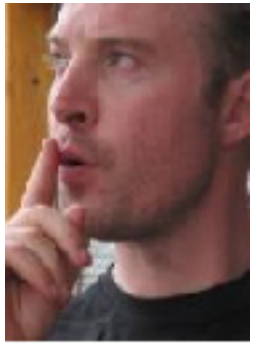
Prof. Werner Schneider
Psychologie

Prof. Katharina Rohlfing
Klinische Linguistik

Dr. Sven Wachsmuth
Imformatik

Prof. David Schlangen
Linguistik

Prof. Thomas Koop
Chemie



Dr. Sander van der Hoog
Wirtschaftswissen-
schaften

Prof. Volker Dürr
Biologie

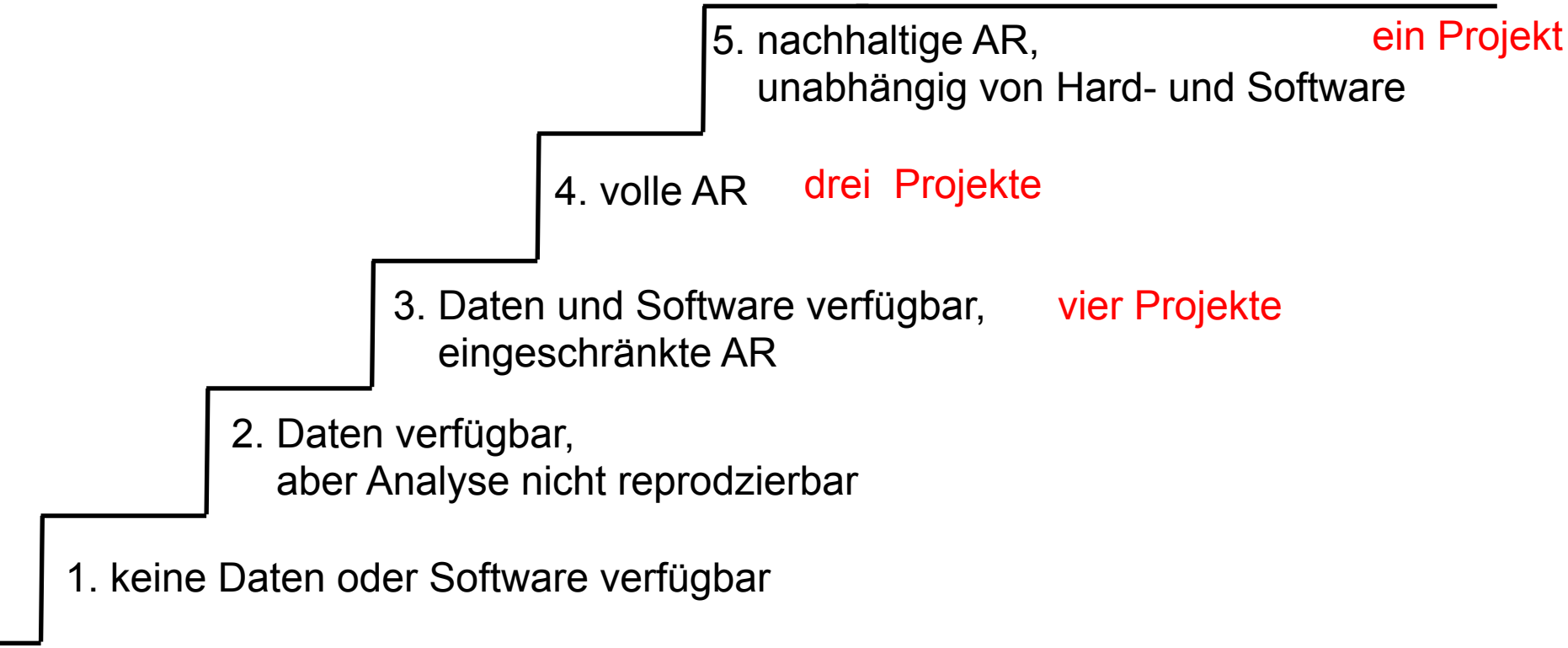
UNIVERSITÄT BIELEFELD

Ergebnisse

- Ergebnisse konnten in allen Fallstudien reproduziert werden
- …allerdings mit erheblichem Aufwand
- …und in vielen Fällen waren zwar Daten und Skripte vorhanden, aber die Dokumentation reichte nicht aus, um die Analysen ohne Schritt-für-Schritt-Anleitung der Autoren*innen der Originalpublikation zu reproduzieren

UNIVERSITÄT BIELEFELD

Analytische Reproduzierbarkeit



UNIVERSITÄT BIELEFELD

Größte Hindernisse

- Fehlen einer Dokumentation und damit die Abhängigkeit von Rücksprache mit den ursprünglichen Autor*innen
- Abhängigkeit von einigen manuellen Schritten im analytischen Arbeitsablauf (z. B. Klicken auf eine GUI)
- Abhängigkeit von nicht-offener und kommerzieller Software
- Fehlen von Informationen darüber, welche Version von Software und/oder Daten verwendet wurde, um ein bestimmtes Ergebnis zu erzeugen.

UNIVERSITÄT BIELEFELD

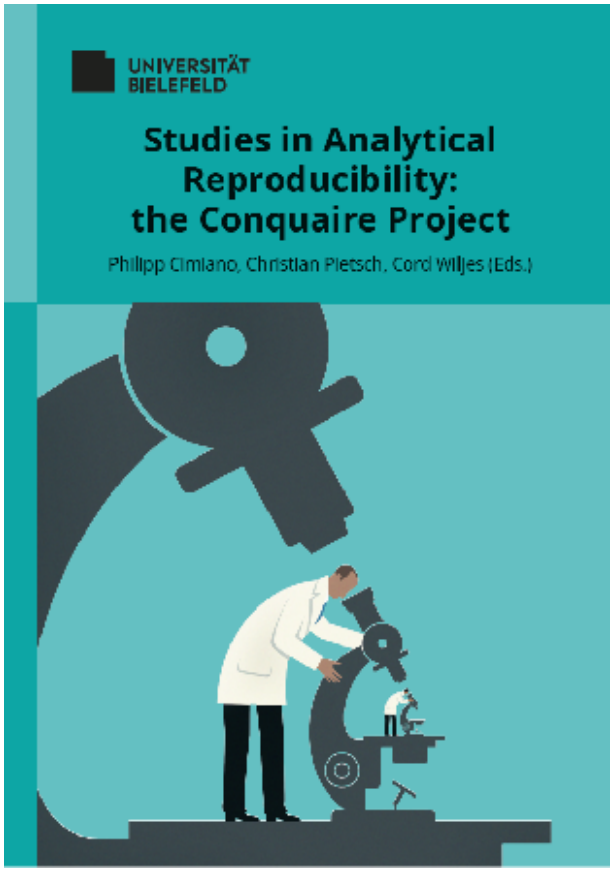
Ideen und weitere Schritte

- Ausbau der Quality Checks
- CI-Workshop/Hackathon

UNIVERSITÄT BIELEFELD

Conquaire Book

Cimiano P., Pietsch C., & Wiljes C. (Eds.) (2021). Studies in Analytical Reproducibility: the Conquaire Project.Bielefeld. DOI: https://doi.org/10.4119/unibi/2942780



UNIVERSITÄT BIELEFELD

Links

- Conquaire Projekt: http://www.uni-bielefeld.de/conquaire/
- Conquaire Project proposal: doi:10.5281/zenodo.31298
- GitLab Mailingliste des DFN: https://www.listserv.dfn.de/sympa/info/gitlab/
- GitLab-AG der Landesinitiative fdm.nrw: https://www.fdm.nrw/index.php/nrw-ag-gitlab/

UNIVERSITÄT BIELEFELD

Vielen Dank!