


CONQUAIRE

Managing data quality in diverse research knowledge bases !

Vidya Ayer
CITEC, Bielefeld University
Germany

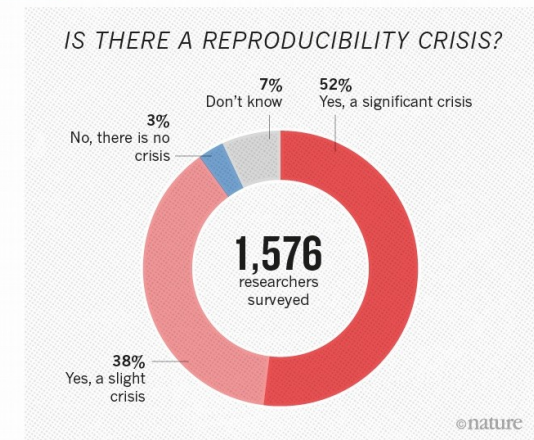
GBV-VZG
Monday, 19 March 2018, Gottingen.
CC BY-NC-SA 4.0 International License.

About

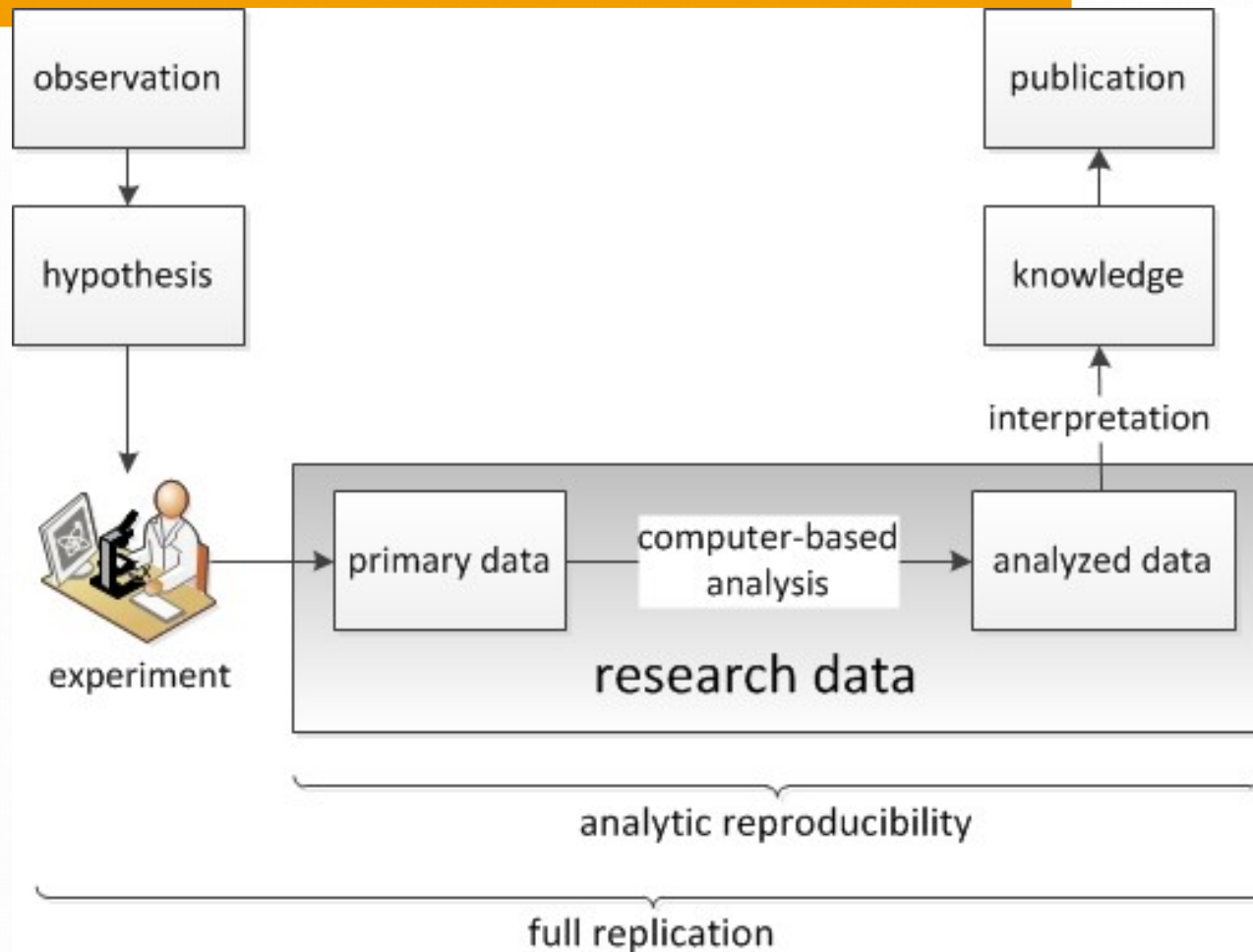
- DFG funded : 2016 – 2019. 
- CITEC + Bielefeld University Library
- Pilot research : 9 partners – interdisciplinary labs

Irreproducibility

- **Reproducibility** – basic principle of Science!
- **Irreproducibility**:
 - Nature survey – scientists (50-70%)
 - Psychology Reproducibility Project (61% fail)
 - Pharma clinical drug trials (82% fail)
- Reproducibility == Full Replication ?



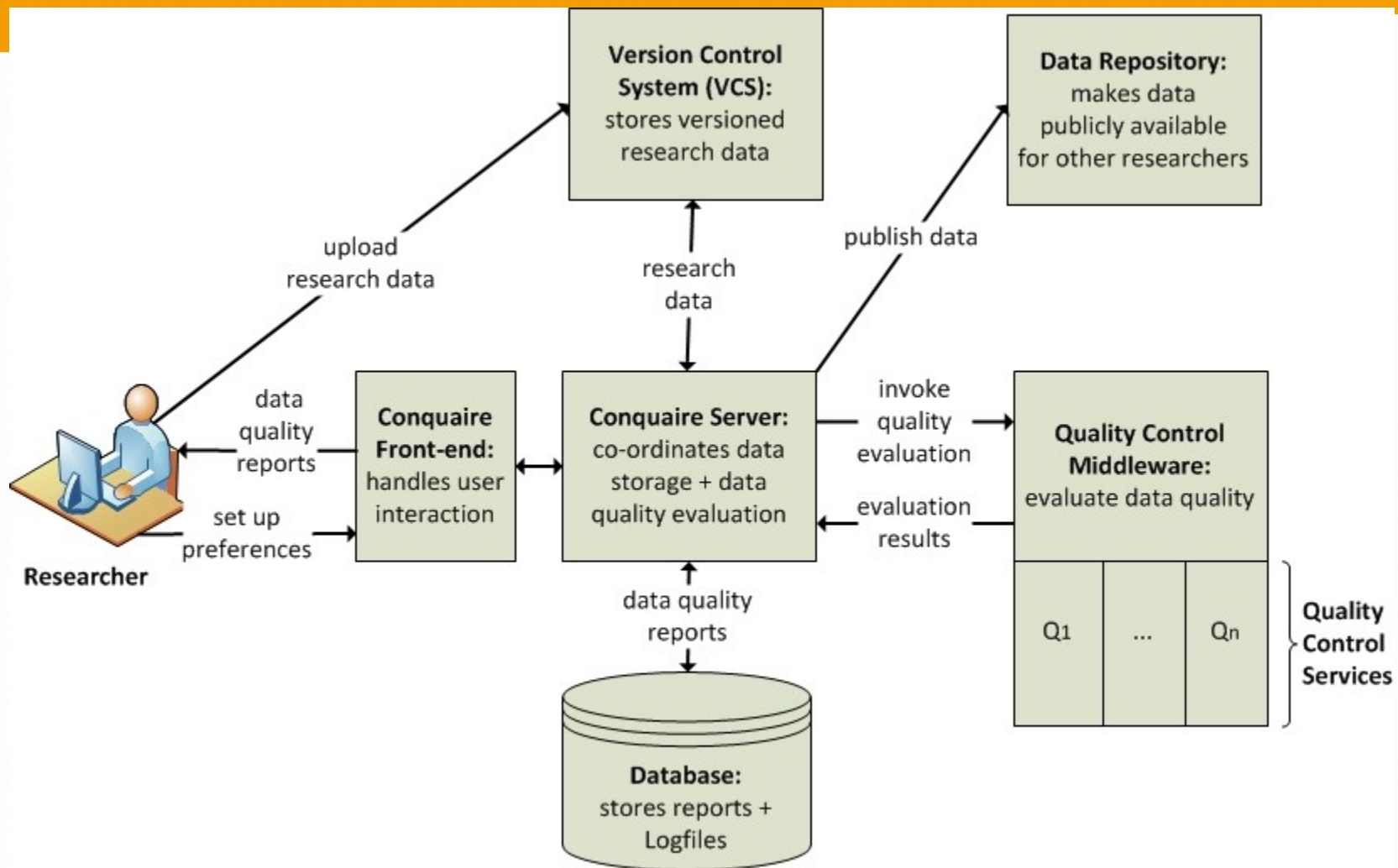
Analytical Reproducibility



Goals

- Research Data Management (RDM) - generic infrastructure
- Support analytical reproducibility of research results
- Storage + DVCS Versioning → data sharing & reuse
- Data Quality = Open formats + Validation (raw, results) + scripts
- Continuous integration → quality data

Architecture



PUB!

Universität Bielefeld

PUB – Publications at Bielefeld University

Deutsch

Publications at Bielefeld University

[Home](#) [Publications](#) [Data Publications](#) [Authors](#) [About PUB](#) [Login](#)

"PUB – Publications at Bielefeld University" is the institutional repository of Bielefeld University. It constitutes the central depot for publication meta data, publications and research data with the mission to reflect the scientific work of the university's researchers. Bielefeld scientists of all disciplines can use this service to create and maintain their personal publication lists. The records are freely accessible and partially linked with fulltexts and research data. [Learn more...](#)

51935 Publication References
Bibliographic data, partially enriched with fulltexts (PDFs etc.)

3427 Individual Author Pages
Documentation of Bielefeld researchers' publishing activities

179 Data Publications
Bibliographic data, partially enriched with research data

9174 Open Access Publications
Freely accessible documents

1853 Theses
Bielefeld dissertations, post-doctoral, and selected master and bachelor theses

Open Access at Bielefeld University

Research Data Management

PUB Theses


Data Seal of Approval

[Contact](#) [Policy](#) [Imprint](#)
Powered by LibreCat
© 2010-2017 Bielefeld University Library

ORCID | Member Organization
Connecting Research and Researchers
Further Memberships & Certificates

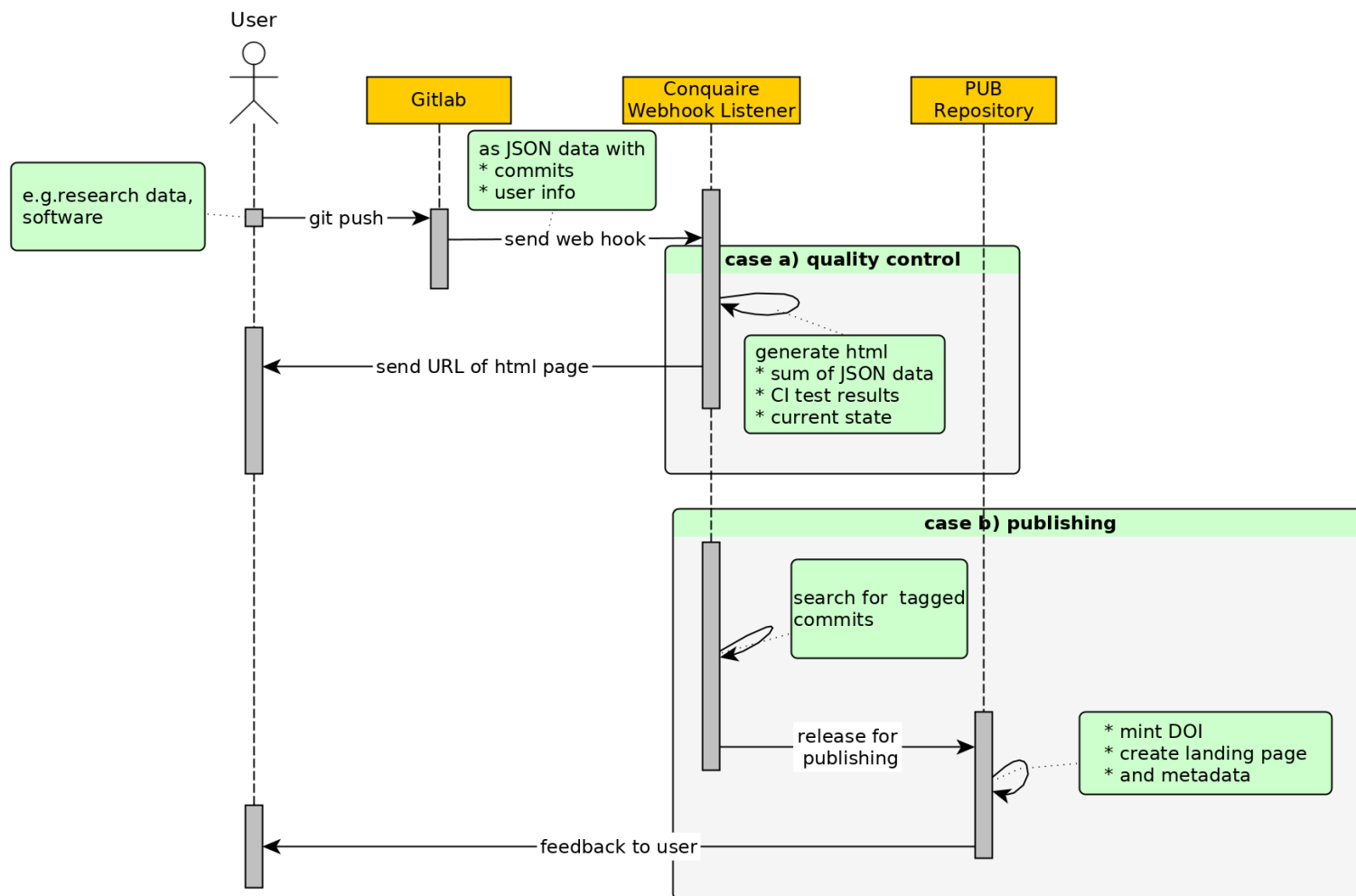
Support Publication Services
Susanne Riedel
Phone: 0521/100-4008
publikationsdienste.ub@uni-bielefeld.de

PUB Overview

- **Management of Institutional research output:**
 - Scientific literature & research data at #UniBi
- **Built with LibreCat:** 
 - Joint effort of Lund, Gent, Bielefeld libraries.
- **Provides:**
 - Author publication lists, supports ORCID
 - Mints DOI / URN for permanent, reliable citation
 - Interfaces (OAI, SRU, CQL)
 - Formats (DC, MODS, DataCite, XMetaDissPlus)
 - Linking of research data with literature
 - Partner in the SAFE Private LOCKSS Network - distributed preservation network.
 - Data Seal of Approval certification



Connecting Gitlab, Conquaire & PUB



DIRA - Data Quality

- Generic quality checks - CSV
 - Ex. declare dtype in format file to process data types.
- Quality checks to support reproducibility & data reusability.
- Diverse file formats:
 - XML, HDF5, JSON, CSV (TSV, Excel sheets with macros), JPEG, MP4, Elan annotated files (.eaf)
- File IO format types issues:
 - '.fdt', '.set', '.mat', '.opj', etc.. require 'non-open' software

Computational reproducibility challenges!

- **Poor data storage solutions**
- **Diverse data formats**
- **FAIR data principles are unfulfilled**
- **Missing data**
- **High maintainence cost [system + (hu)manpower]**
- **Manual file handling of research data – error prone**
- **Unclean datasets**
- **Data analysis pipeline not fully automated**

Rainbow-hued Research Data

- **Libraries = Knowledge Base (BigData + DataScience)**
 - Not only preserving and archiving data
 - Analyse research data
 - Extract knowledge from it
 - Efficient data storage management
 - Key to discover solutions to multitude of different problems.

Dark (Research) Data

- **Analysis challenges for libraries**
 - No documentation of data (or the analysis process)
 - Dump of data files when research ends
 - No Metadata labels or ontologies – semantic data
 - No Open Datasets – no data interoperability
 - Unstructured data - raw (video or image) data analysis?
 - Machine learning solutions?
 - Training and support – dark data analysis?

Thank You!

- Questions?
- Contact:
 - Email: **ayer@uni-bielefeld.de**
 - Github & Twitter: **@svaksha**
 - Email (project):
conquaire-contact@lists.uni-bielefeld.de
 - Website: **<http://conquaire.uni-bielefeld.de>**