Universität Bielefeld

CITEC

CONQUAIRE

Towards an architecture supporting continuous quality control to ensure reproducibility of research.

Vidya Ayer, Christian Pietsch, Johanna Vompras,
Jochen Schirrwagen, Najko Jahn, Cord Wiljes, Philipp Cimiano.
CITEC, Bielefeld University, Germany

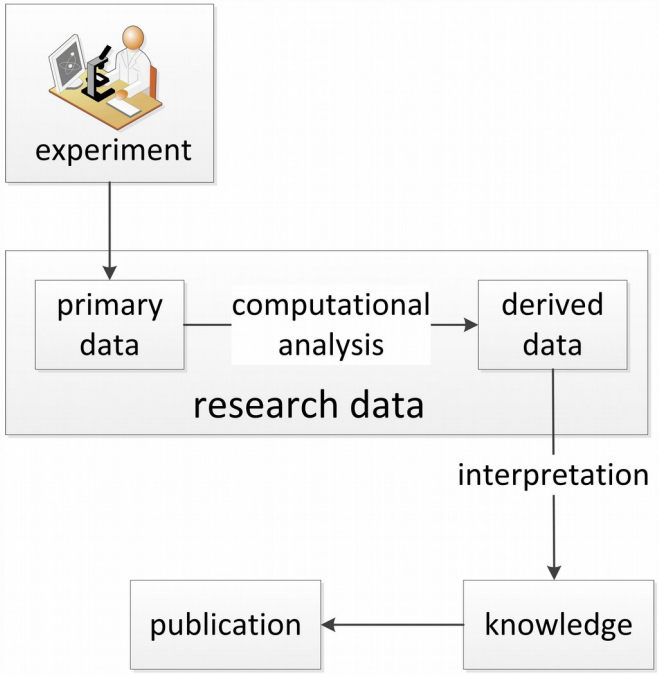
Reproducible Open Science Workshop
Hannover, Germany, 2016-09-09

OUTLINE

- Introduction
- Irreproducibility
- Objectives
- Data Desiderata
- Case Studies
- Research Lifecycle
- VCS : Advantages Vs. Disadvantages
- Architecture
- Quality Control
- Conclusion
- ThankYou! (Questions? / Contact)

INTRODUCTION

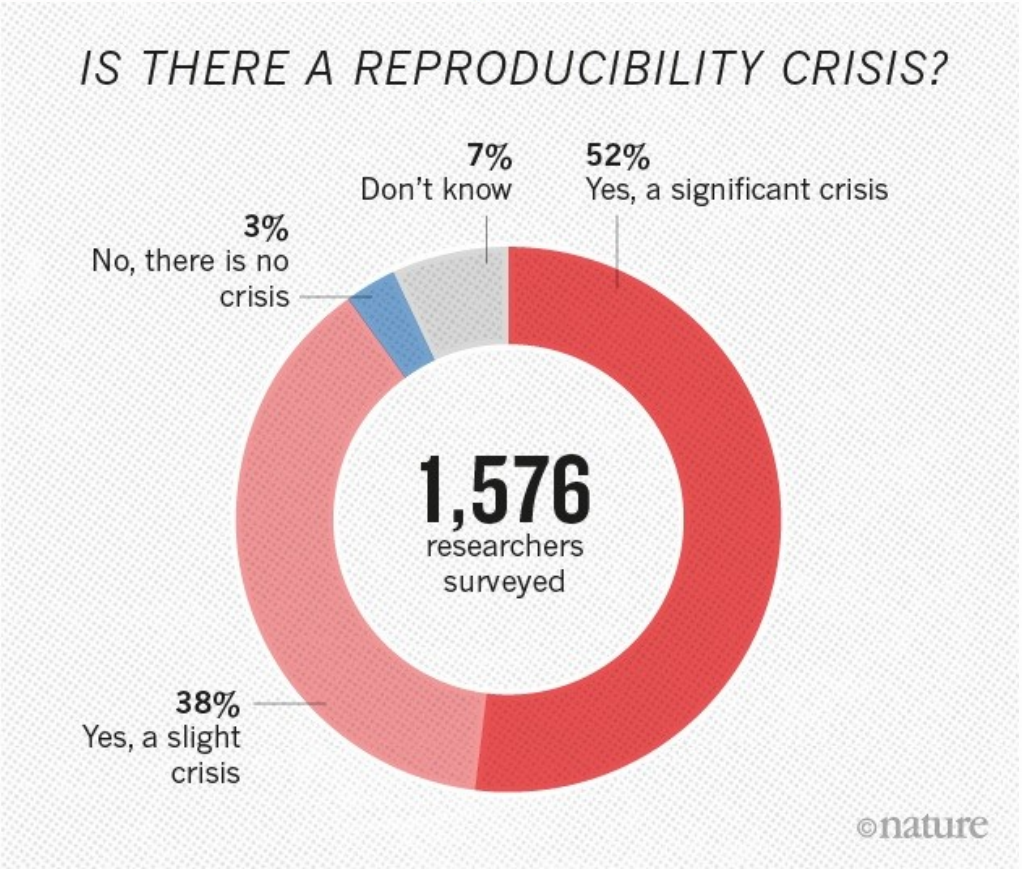
- DFG funded
  - February 2016 – January 2019
- CITEC + Bielefeld University Library



- Replication Vs. Reproducibility
- Analytical Reproducibility = Mathematical!
- Continuous quality control

IRREPRODUCIBILITY

- Metastudies
- Computer Science, Psychology, Medicine
- Failure rates :
  - 70% others research
  - > 50% own research
  - < 31% failure = wrong result



Credits: Baker M., Is there a reproducibility crisis? Nature, 2016.

OBJECTIVES

- Infrastructure - research reproducibility
- Research Data Management System (RDMS)
- Storage + Versioning => Not Enough
- Data = quality + open formats
- Data Validation = raw + results
- Computational analytical reproducibility
- Manage scripts

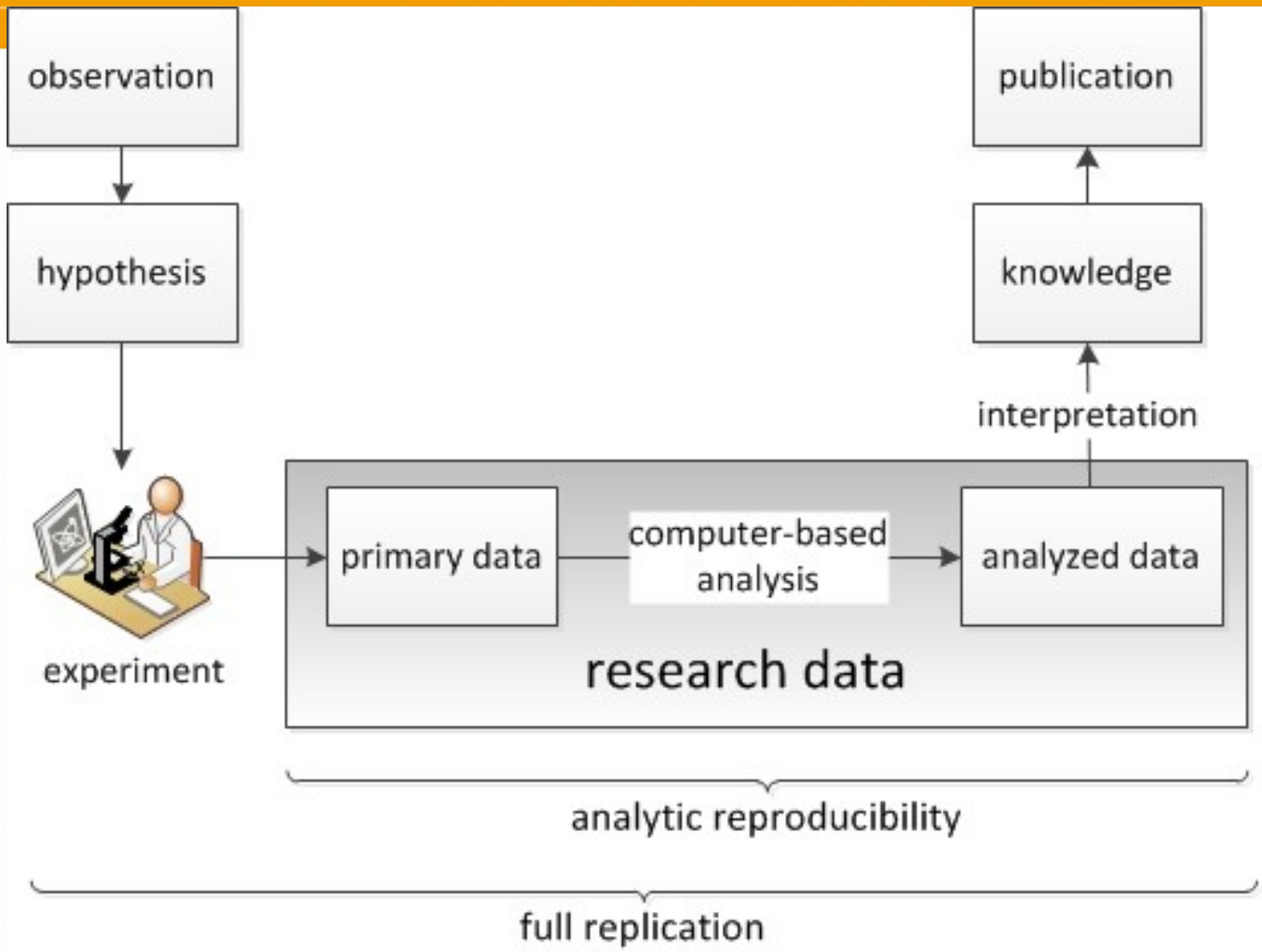
DATA DESIDERATA

- Common format research data : download / inspect
- Syntactically valid open format data
- Documentation : data, scripts, experiment condition
- Third party researcher : data element semantics / understand / reproduce
- Analytical procedures - data processing
- Independent researcher / analyst
  - re-run analytical pipeline
  - verify published results
  - Free and Open Source Software (FOSS)

CASE STUDIES

- **Pilot project**:
  - 9 research partners
  - Interdisciplinary + InterUniversity
- **Disciplines**: Applied Computational Linguistics, Biology, Computer Science, Chemistry, Economics, Linguistics, Neurobiology, Psychology, Sports Science
- Varied data formats/ experimental tools/ software
- Data Management Plan (DMP)

RESEARCH LIFECYCLE



VCS
Advantages Vs. Disadvantages

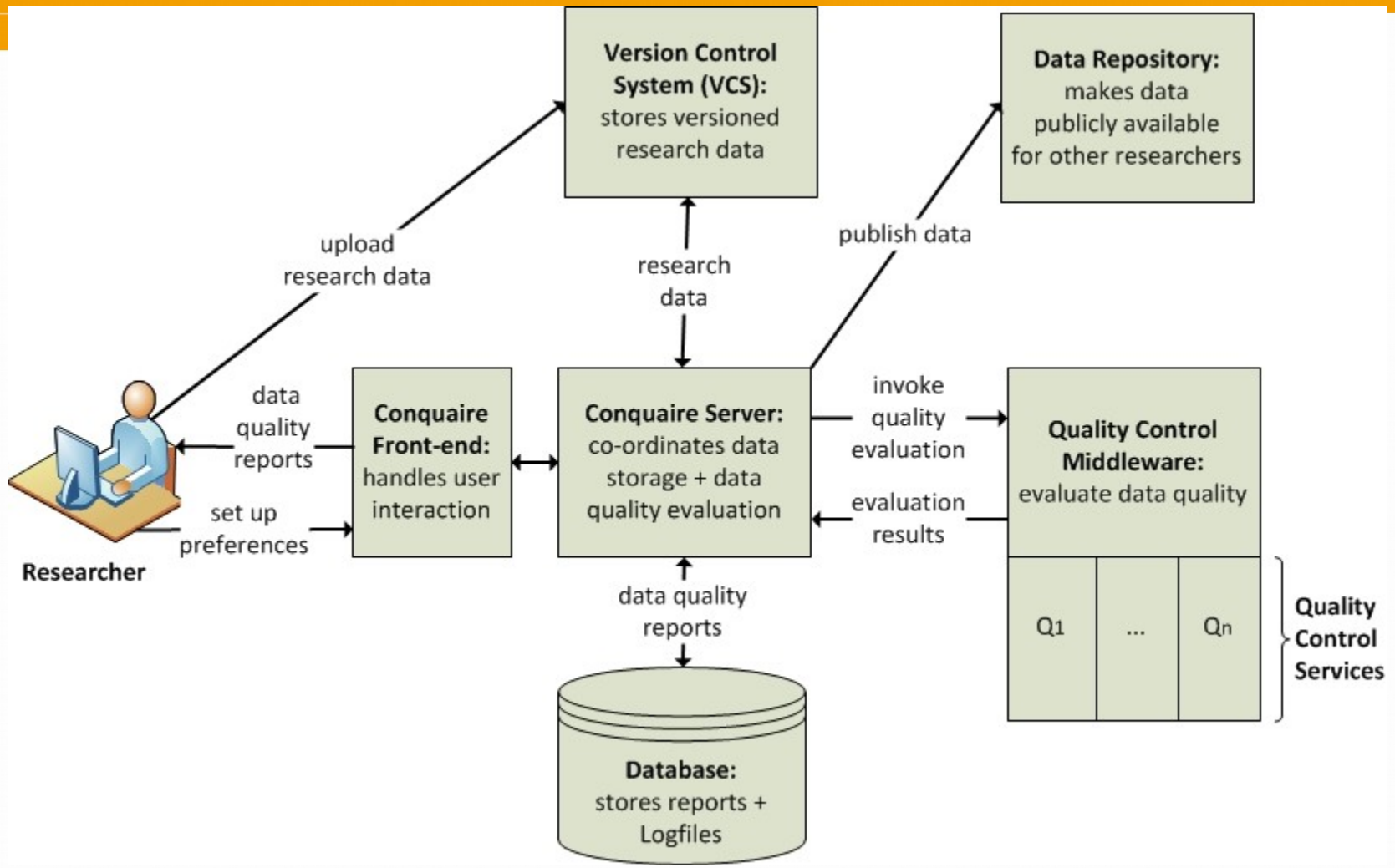
- Solves data tracking problem
- Access to data revisions
- Timestamps : data + scripts
- Commit : Revision, Diffs, Merges, Branch.
- Objects : compressed, SHA1 hash
- Stored as 'pack-file'
- CI (runners), daemons, loggers

- 3-5 years : data loss
- No support for change tracking
- Manual file comparison
- Data sharing challenges
- Dropbox, Sciebo, Google Drive, private servers/ disks
- Cloud services => Privacy concerns, no backups
- Data changes over time

ARCHITECTURE

- Front-End GUI : repo status, badges, analysis reports
- Back-End Architecture :
  - VCS Server : research artefacts
  - Conquaire Server : Stateful Process Logic
  - Quality Analysis Middleware Server
    - Data management / curation
    - Continuous Integration (CI)
    - Quality analysis / processing
    - Monitor project data / metadata
    - Standards : ISO, IETF, W3C
  - Database
- Front-End : PUB (Publications at Bielefeld University) paper downloads
  - Trusted Data Repository integration (Back-End)

ARCHITECTURE



QUALITY CONTROL

- Analysis : compute statistical functions
- Documentation, Scripts, Primary/Secondary data (raw + results)
- Support multiple data exchange formats
- **.csv** sample implementation
  - Metadata schema standard : YAML format

- Standards specification : **RFC 4180**
  - data type matching
  - records (one per line)
  - column headers
  - range of values : metadata
  - field delimiters (reserved character: comma, semicolon, or tab)
  - column textual definition : warn < 5 tokens
  - out of range value / NAN / Null
  - character set such as ASCII, various Unicode character sets (e.g. UTF-8), EBCDIC

CONCLUSION

- Conquaire : Infrastructure for research data management system (RDMS)
  - Support continuous data quality control
  - Enable analytical reproducibility

- Collaborators – Bielefeld University research partners
  - Store research artefacts (data, analytical workflows, software, publications)
  - Improve research methods and workflow tools

- Current Architechture status
  - Created pre-alpha implementation
  - Quality System functionality – ex. CSV

THANK YOU!

- Questions?

- Contact Information:
  - Email: vayer@techfak.uni-bielefeld.de
  - Project Email: conquaire-contact@lists.uni-bielefeld.de
  - Website: http://conquaire.uni-bielefeld.de